UNIVERSITÀ DEGLI STUDI MILANO BICOCCA

FACOLTÀ DI ECONOMIA

CORSO DI LAUREA IN MARKETING, COMUNICAZIONE AZIENDALE E MERCATI GLOBALI

LA "CLUSTER ANALYSIS": UN'APPLICAZIONE AL MERCATO AUTOMOBILISTICO

TESI DI LAUREA DI

STEFANO GIUSEPPE PANDINI

MATRICOLA 726713

RELATORE: PROF. ALESSANDRO ZINI

ANNO ACCADEMICO 2012 - 2013

Indice:

Introduzione

La ricerca che mi propongo di effettuare è un analisi cluster del settore automobilistico italiano. L'obbiettivo è suddividere i modelli di auto disponibili sul mercato in gruppi i più omogenei possibili al loro interno.

Per effettuare l'analisi è stato usato un dataset proveniente dalla rivista mensile Quattruote (Editoriale Domus). L'analisi è stata svolta tramite il software statistico SPSS (IBM).

La tesi è stata scritta con metodo scientifico, supponendo che il lettore sia già a conoscenza delle tecniche usate durante l'analisi e non abbia bisogno di spiegazioni. Nell'elaborato quindi non saranno presenti le definizioni delle statistiche e delle procedure usate, ma solo le motivazioni che hanno portato ad una determinata scelta e le conclusioni a cui si è arrivati.

Di tutti i mercati su cui era possibile svolgere questa analisi, è stato scelto il mercato automobilistico per via della sua importanza: non solo rappresenta ben l'11% del PIL italiano, ma è anche il settore industriale con cui i consumatori entrano in contatto più frequentemente. Ancora nel 2014, più di un secolo dopo la sua invenzione, l'automobile resta il mezzo di trasporto più comune e la sua importanza non accenna a diminuire. Data la sua natura rivolta al consumatore, il settore automobilistico è uno dei settori più colpiti dalla crisi. È anche però il settore a cui si guarda sempre in cerca di un segnali positivi: ci sono infatti pochi indicatori di una ripresa dei consumi più veritieri del numero di immatricolazioni delle auto.

Una buona classificazione di questo mercato può avere molte applicazioni: comprendere meglio il posizionamento dei propri veicoli, conoscere la concorrenza con cui devono gareggiare, trovare i segmenti più proficui e quelli più facilmente penetrabili.

Cercando di evitare la banale classificazione delle automobili in base al rapporto prezzo/prestazioni sono state tralasciate nell'analisi molte variabili tecniche, come i cavalli e l'accelerazione, in favore di alcune variabili che negli ultimi anni hanno acquistato sempre più importanza, come le emissioni e i consumi.

Queste variabili hanno andamento contrario rispetto a quelle sulle prestazioni (più un auto va veloce meglio è, più consuma un auto più è caro mantenerla) e si ritiene che sempre più consumatori ne tengano conto al momento dell'acquisto di una nuova auto.

La tesi è strutturata in questo modo:

- nel secondo capitolo si è proceduto a controllare i dati per eventuali errori ed a ridurli per agevolare il calcolo.

- Nel terzo capitolo è stato necessario estrarre le componenti principali dalle variabili, in modo da eliminarne la correlazione.

- Nel quarto capitolo si è svolta l'analisi vera e propria: i casi sono stati raggruppati tramite un algorimo di classificazione gerarchico.

- Nel quinto capitolo, si è passati all'analisi esplorativa dei cluster trovati.

Nelle conclusioni si è infine cercato di capire se la procedura utilizzata ha avuto successo e se i cluster trovati rappresentano una buona classificazione.

Costruzione del dataset

Il dataset su cui è stata svolta la nostra analisi è stato estratto dall'edizione di Gennaio 2014 della rivista mensile Quattroruote.

Inizialmente il nostro dataset comprendeva 9 variabili e oltre 4100 casi, è stato dunque necessario ridurlo sia per problemi di calcolo (un normale computer non è in grado l'analisi su una matrice così ampia) sia per rendere l'analisi più chiara e agevolare l'interpretazione. Il dataset è stato ridotto sia in lunghezza che in larghezza.

Delle variabili si è deciso di tenere quelle che si ritiene sianno più importanti per i consumatori, ovvero:

- **Prezzo:** variabile numerica espressa ovviamente in Euro. Va da un minimo di 7.750 per la Renault - Twingo 1.2 ad un massimo di 365.000 per la Lamborghini - Aventador.
- **Cilindrata:** variabile numerica espressa in cm^3. Va un minimo di 799 per la Smart ForTwo 800 ad un massimo di 6498 per la Lamborghini - Aventador. La cilindrata del motore è un concetto non applicabile ai motori elettrici, quindi alle auto elettriche è stato assegnato il valore 0.
- **Alimentazione:** variabile testuale su scala nominale. È presente in 6 modalità, le più importanti sono: "Benzina", il 45% del totale, e "Diesel", il 46% del totale.
- **Emissioni:** variabile numerica espressa in grammi di CO_2 al Kilometro. Va da un minimo di 0 per le auto elettriche ad un massimo di 398 sempre per la Lamborghini - Aventador.
- **Velocità massima:** variabile numerica espressa in Km/h. Va da un minimo di 130 per la Citroen - C-Zero ad un massimo di 350 per Lamborghini - Aventador.
- **Consumo misto:** variabile numerica espressa in litri di carburante (oppure Kg di Metano) per 100 km. Va da un minimo di 0 per le auto elettriche ad un massimo di 17.2 per la Lamborghini - Aventador.

Per poter utilizzare la variabile alimentazione in una cluster gerarchica, una discriminante troppo importante ormai nella scelta di una nuova automobile per essere tralasciata, è stato necessario trasformarla in una variabile numerica in questo modo:

Codice	Descrizione	Variabile Numerica
B	BENZINA	1
BG	BENZINA/GPL	2
BM	BENZINA/METANO	3
D	DIESEL	4
E	ELETTRICA	5
I	IBRIDA	6

L'ordine in cui sono stati inseriti i vari tipi di alimentazione è quello indicato nell'indice della rivista Quattroruote.

Sono stati eseguiti dei controlli per verificare la presenza di errori nel dataset tramite campioni casuali e agli estremi delle distribuzioni.

Per quanto riguarda la riduzione dei casi si è proceduto in una prima selezione ad eliminare tutte le versioni di modelli auto per cui, oltre al prezzo, non variava nessun altra variabile.

Osservando però che dopo questa scrematura rimanevano ancora più di 3000 modelli di auto si è proceduto con una seconda, più drastica, selezione.

In questa fase si è deciso di eliminare tutte le versioni di un automobile che non fossero quella di base, come possono essere le versioni "sport", "chic" e "elegance". Sono invece state tenute le diverse versioni di cilindrata dei modelli, in quanto presentano differenze molto significative e richiamano consumatori molti diversi: un pubblico più adulto e maschile per le cilindrate alte, uno più giovane per quelle basse.

Quando si farà riferimento quindi ad un determinato modello nella tesi quindi si indicherà la cilindrata a cui ci si riferisce, intendendo però sempre la versione di base del modello, ovvero quella acquistabile al minor prezzo.

Potete vedere qui sotto un esempio della selezione effettuata nelle due fasi.

Eliminati nella prima selezione

Marca - Modello	Prezzo	Cilindrata	Alimentaz.	Emissioni	Velocità	Consumo misto
Citroen - C3 1.0 VTi 68 Attraction	12.650	999	B	99	155	4,3
Citroen - C3 1.0 VTi 68 Seduction	14.150	999	B	99	155	4,3
Citroen - C3 1.2 e-VTi 82 ETG air.	15.900	1199	B	95	176	4,1
Citroen - C3 1.2 e-VTi 82 ETG air. Vanity Fair 10	16.400	1199	B	99	176	4,3
Citroen - C3 1.2 e-VTi 82 ETG airdream Exclusive	17.400	1199	B	99	176	4,3
Citroen - C3 1.2 VTi 82 Exclusive	16.650	1199	B	109	174	4,7
Citroen - C3 1.2 VTi 82 Seduction	14.900	1199	B	109	174	4,7
Citroen - C3 1.4 e-HDi 70 airdream CMP Seduction	17.150	1398	D	87	165	3,4
Citroen - C3 1.4 HDi 70 Exclusive	17.900	1398	D	99	163	3,8
Citroen - C3 1.4 HDi 70 Seduction	16.150	1398	D	99	163	3,8
Citroen - C3 1.4 HDi 70 Vanity Fair 10	16.900	1398	D	99	163	3,8
Citroen - C3 1.4 VTi 95 GPL airdream Exclusive	18.400	1397	BG	129	182	8,2
Citroen - C3 1.4 VTi 95 GPL airdream Seduction	16.650	1397	BG	127	184	8,1
Citroen - C3 1.6 e-HDi 115 airdream Exclusive	19.650	1560	D	99	190	3,8
Citroen - C3 1.6 e-HDi 90 airdream Exclusive	18.900	1560	D	90	180	3,5

Eliminati nella seconda selezione

Marca - Modello	Prezzo	Cilindrata	Alimentaz.	Emissioni	Velocità	Consumo Misto
Citroen - C3 1.0 VTi 68 Attraction	12.650	999	B	99	155	4,3
Citroen - C3 1.2 e-VTi 82 ETG air.	15.900	1199	B	95	176	4,1
Citroen - C3 1.2 e-VTi 82 ETG air. Vanity Fair 10	16.400	1199	B	99	176	4,3
Citroen - C3 1.2 VTi 82 Exclusive	16.650	1199	B	109	174	4,7
Citroen - C3 1.4 e-HDi 70 airdream CMP Seduction	17.150	1398	D	87	165	3,4
Citroen - C3 1.4 HDi 70 Seduction	16.150	1398	D	99	163	3,8
Citroen - C3 1.4 VTi 95 GPL airdream Exclusive	18.400	1397	BG	129	182	8,2
Citroen - C3 1.4 VTi 95 GPL airdream Seduction	16.650	1397	BG	127	184	8,1
Citroen - C3 1.6 e-HDi 115 airdream Exclusive	19.650	1560	D	99	190	3,8
Citroen - C3 1.6 e-HDi 90 airdream Exclusive	18.900	1560	D	90	180	3,5

Risultato finale

Marca - Modello	Prezzo	Cilindrata	Alimentaz.	Emissioni	Velocità	Consumo Misto
Citroen - C3 1.0 VTi 68 Attraction	12.650	999	B	99	155	4,3
Citroen - C3 1.2 e-VTi 82 ETG air.	15.900	1199	B	95	176	4,1
Citroen - C3 1.4 HDi 70 Seduction	16.150	1398	D	99	163	3,8
Citroen - C3 1.4 VTi 95 GPL	16.650	1397	BG	127	184	8,1
Citroen - C3 1.6 e-HDi 90	18.900	1560	D	90	180	3,5

Dei modelli rimanenti, uno solo è stato escluso arbitrariamente dall'analisi in quanto così particolare da essere in grado di alterare da solo i nostri risultati. Il modello in questione è la Mercedes - SLS AMG Electric Drive Coupé, un'automobile unica e che non ha semplicemente alcun concorrente sul mercato. È in assoluto l'automobile più costosa di tutte, con prestazione da top di gamma e le caratteristiche di una comune auto elettrica.

Marca - Modello	Prezzo	Cilindrata	Alimentaz.	Emissioni	Velocità	Consumo M.
Mercedes - SLS AMG Electric	432.000	0	E	0	250	0

I rimanenti 1396 modelli rappresentano il dataset finale, sul quale si è svolta l'analisi.

L'analisi delle componenti principali

Il dataset che si è costruito nel capitolo precedente presenta un difetto che potrebbe compromettere la nostra analisi cluster: le variabili che abbiamo considerato sono fortemente correlate. In particolare le variabili prezzo, cilindrata, emissioni, velocità e consumi hanno andamenti noti e, volendo, si può prevedere molto bene una di esse in funzione delle altre.

L'alta correlazione tra le variabili è un problema, perché crea ridondanze nei dati che vengono contate nel processo di raggruppamento, distorcendo i risultati.

Per dare un'idea della correlazione tra le variabili, si noti la seguente matrice di correlazione:

Matrice di correlazione

	Prezzo	Cilindrata	Alimentazione	Emissioni	VelocitàMax	ConsumoMisto
Prezzo	1,000	,850	-,094	,747	,781	,693
Cilindrata	,850	1,000	-,057	,836	,779	,761
Alimentazione	-,094	-,057	1,000	-,356	-,174	-,483
Emissioni	,747	,836	-,356	1,000	,677	,951
VelocitàMax	,781	,779	-,174	,677	1,000	,621
ConsumoMisto	,693	,761	-,483	,951	,621	1,000

Più i valori sono vicini a 1 e -1 più le variabili sono correlate. È facile notare come quasi tutte le variabili, eccetto l'alimentazione, siano correlate con le altre.

Per far sì che la correlazione non influenzi l'analisi solitamente viene usata nell'analisi la distanza di Mahalanbis, in quanto più adeguata quando si hanno variabili correlate. Purtroppo questa distanza non è disponibile nel software SPSS e nemmeno di facile implementazione.

Per ovviare a questa mancanza si è scelto di effettuare un'analisi delle componenti principali prima di svolgere l'analisi cluster.

Con la PCA vengono estratte dalle variabili le componenti principali, che sono incorrelate tra di loro ed eliminano così il nostro problema.

Nell'analisi PCA che è stata eseguita si è scelto di estrarre tutte le componenti principali con autovalori maggiori di 1, il valore standard in grado estrarre tutte le componenti principali significative per la nostra analisi.

Non si tiene conto di quelle con autovalori minori di 1 perché, non essendoci modo per pesare le variabili con il software SPSS, se le lasciassimo finiremmo dare a componenti principali di modesto valore lo stesso peso delle componenti principali significative.

La soluzione trovata non è stata soggetta a nessuna rotazione, in quanto fine dell'elaborato è un'analisi cluster e non interpretare i fattori estratti.

L'analisi svolta con il software SPSS ha estratto 2componenti principali.

Test KMO (Keiser Meyer Olkin) e di Bartlett

Misura di adeguatezza campionaria KMO		,792
Test di sfericità di Bartlett	Chi-quadrato appross.	9276,904
	df	15
	Sig.	,000

La misura KMO indica quanta parte di varianza è spiegata da fattori comuni. Il test di sfericità di Bartlett testa l'ipotesi che la nostra matrice di correlazione sia una matrice identità, il che indicherebbe che le nostre variabili sono incorrelate. Perchè i dati siano adatti all'analisi fattoriale la misura KMO dev'essere maggiore di 0.7 e la significatività del test di Bartlett dev'essere minore di 0.05. Possiamo procedere con l'analisi PCA.

Comunalità

	Iniziale	Estrazione
Prezzo	1,000	,858
Cilindrata	1,000	,920
Alimentazione	1,000	,940
Emissioni	1,000	,902
VelocitàMax	1,000	,749
ConsumoMisto	1,000	,907

Metodo di estrazione: Analisi componenti principali.

Anche dopo l'estrazione i valori di comunalità sono alti per tutte le variabili.

Varianza totale spiegata

Componente	Autovalori iniziali			Pesi dei fattori non ruotati		
	Totale	% di varianza	% cumulata	Totale	% di varianza	% cumulata
1	4,169	69,490	69,490	4,169	69,490	69,490
2	1,107	18,449	87,939	1,107	18,449	87,939
3	,397	6,625	94,564			
4	,187	3,115	97,678			
5	,103	1,720	99,399			
6	,036	,601	100,000			

Con 2 componenti principali estratte la varianza comulata spiegata è il 87,4%, una percentuale decisamente buona.

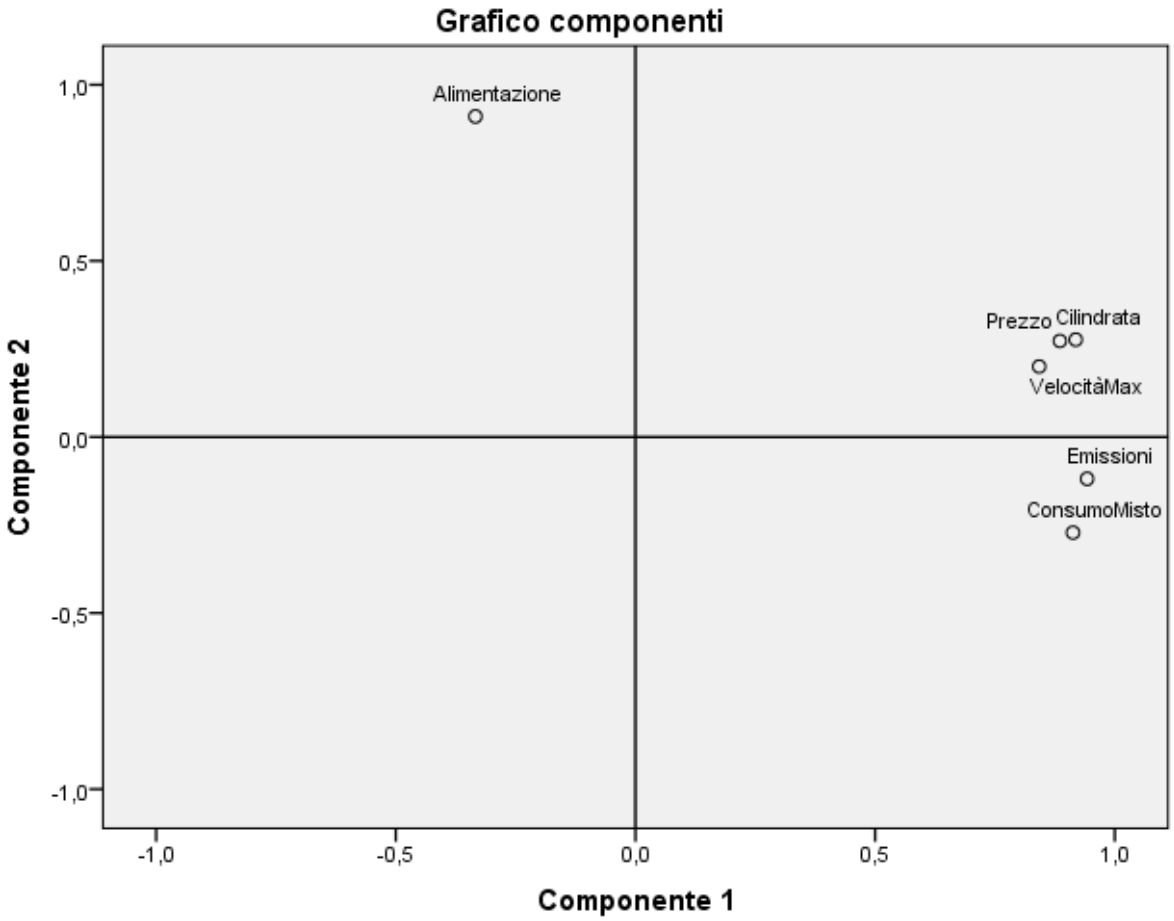
Matrice di componenti[a]

	Componente	
	1	2
Emissioni	,942	-,119
Cilindrata	,918	,276
ConsumoMisto	,913	-,272
Prezzo	,885	,273
VelocitàMax	,842	,200
Alimentazione	-,334	,910

Metodo estrazione: analisi componenti principali.

a. 2 componenti estratti

Più i valori sono vicini a 1 più la componente principale è associata ad una variabile. La prima componente principale è associata alle variabili Emissioni, Cilindrata, Consumo, Prezzo e Alimentazione. La seconda componente principale è corrisponde praticamente solo all'Alimentazione.



Le due componenti principali trovate sono state salvate come variabili, sulle quali è stata eseguita la cluster analisi.

L'analisi cluster

Dopo numerosi tentativi di prova per vedere quale metolodogia restituiva risultati migliori è stato scelto come metodo di aggregazione il metodo di Ward. È un metodo molto usato perché fornisce risultati equilibrati e può essere usato con qualunque distanza.

Potendo scegliere qualunque distanza, nell'analisi si è scelto di utilizzare come distanza di aggregazione quella consigliata dal software SPSS per il metodo di Ward, ovvero la distanza euclidea quadratica.

Data l'elevata numerosità in esame non è possibile mostrare l'output completo del software. Verranno quindi inseriti nelle prossime pagine solo gli estratti più importanti dell'analisi, lasciando la possibilità di consultare i risultati completi nell'appendice A.

La tabella di aggregazione risultante dalla nostra analisi è la seguente.

Programma di agglomerazione

Stadio	Cluster accorpati		Coefficienti	Stadio di formazione del cluster		Stadio successivo
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	717	718	,000	0	0	313
2	498	499	,000	0	0	20
3	271	274	,000	0	0	85
4	47	214	,000	0	0	189
6	502	550	,000	0	0	67
...	...	...	...	...	...	...
1382	958	980	146,478	1355	1352	1391
1383	739	868	156,695	1369	1379	1387
1384	867	903	173,743	1361	1374	1391
1385	2	7	191,739	1378	1348	1392
1386	449	503	215,260	1297	1370	1390
1387	739	875	246,941	1383	1372	1393
1388	433	668	299,914	1377	1380	1390
1389	1	381	353,702	1381	1376	1392
1390	433	449	415,326	1388	1386	1393
1391	867	958	481,390	1384	1382	1394
1392	1	2	650,742	1389	1385	1394
1393	433	739	825,056	1390	1387	1395
1394	1	867	1552,738	1392	1391	1395
1395	1	433	2790,000	1394	1393	0

La colonna “coefficienti” indica la distanza a cui vengono uniti i due Cluster. Le altre colonne hanno significati elementari.

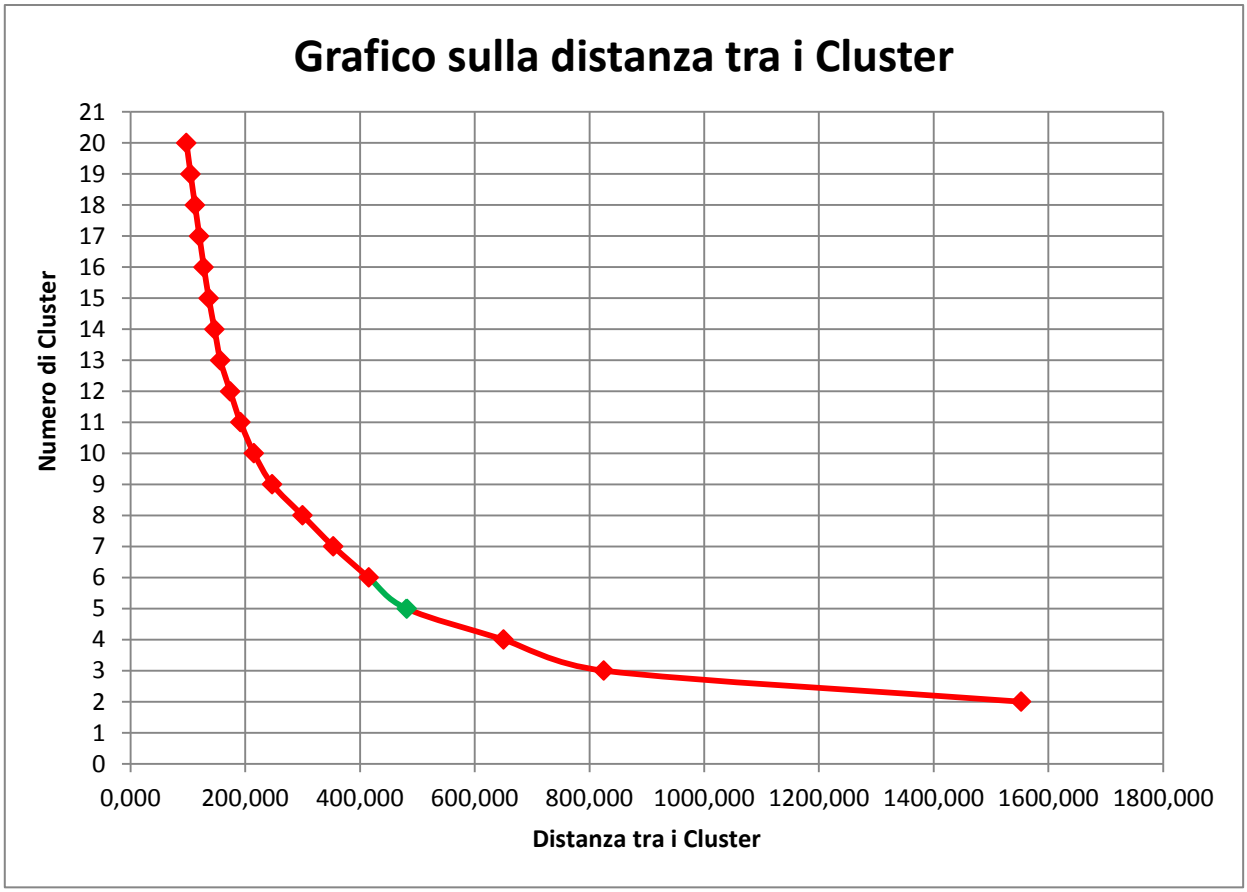
Utilizzando le ultime 20 righe della tabella di aggregazione è stata creata la seguente tabella.

Stadio	Distanza	Incremento	Incremento percentuale	N° Cluster
1376	97,564	6,467	0,42%	20
1377	104,730	7,166	7,34%	19
1378	112,358	7,628	7,28%	18
1379	120,080	7,722	6,87%	17
1380	127,933	7,854	6,54%	16
1381	136,486	8,553	6,69%	15
1382	146,478	9,992	7,32%	14
1383	156,695	10,217	6,98%	13
1384	173,743	17,047	10,88%	12
1385	191,739	17,996	10,36%	11
1386	215,260	23,521	12,27%	10
1387	246,941	31,681	14,72%	9
1388	299,914	52,973	21,45%	8
1389	353,702	53,787	17,93%	7
1390	415,326	61,624	17,42%	6
1391	481,390	66,064	15,91%	5
1392	650,742	169,352	35,18%	4
1393	825,056	174,314	26,79%	3
1394	1552,738	727,681	88,20%	2
1395	2790,000	1237,262	79,68%	1

Il numero ottimale di cluster si trova guardando l’incremento maggiore tra le varie distanze di aggregazione. Una volta trovato il passo a cui avviene il “salto”, lo si sottrae al numero totale di casi e si ottiene il numero di ottimo di raggruppamenti.

In questo caso, il maggior incremento è chiaramente tra il terzultimo e il penultimo raggruppamento: un incremento del 88.2% rispetto al 26.79% del passo precedente. Questa soluzione non è però quella migliore, in quanto dobbiamo scegliere la soluzione per la quale è più evidente la discontinuità tra i gruppi: la cosiddetta soluzione “a gomito”, dove la distanza tra i cluster aumenta abbastanza da rendere la spezzata delle distanze più orizzontale.

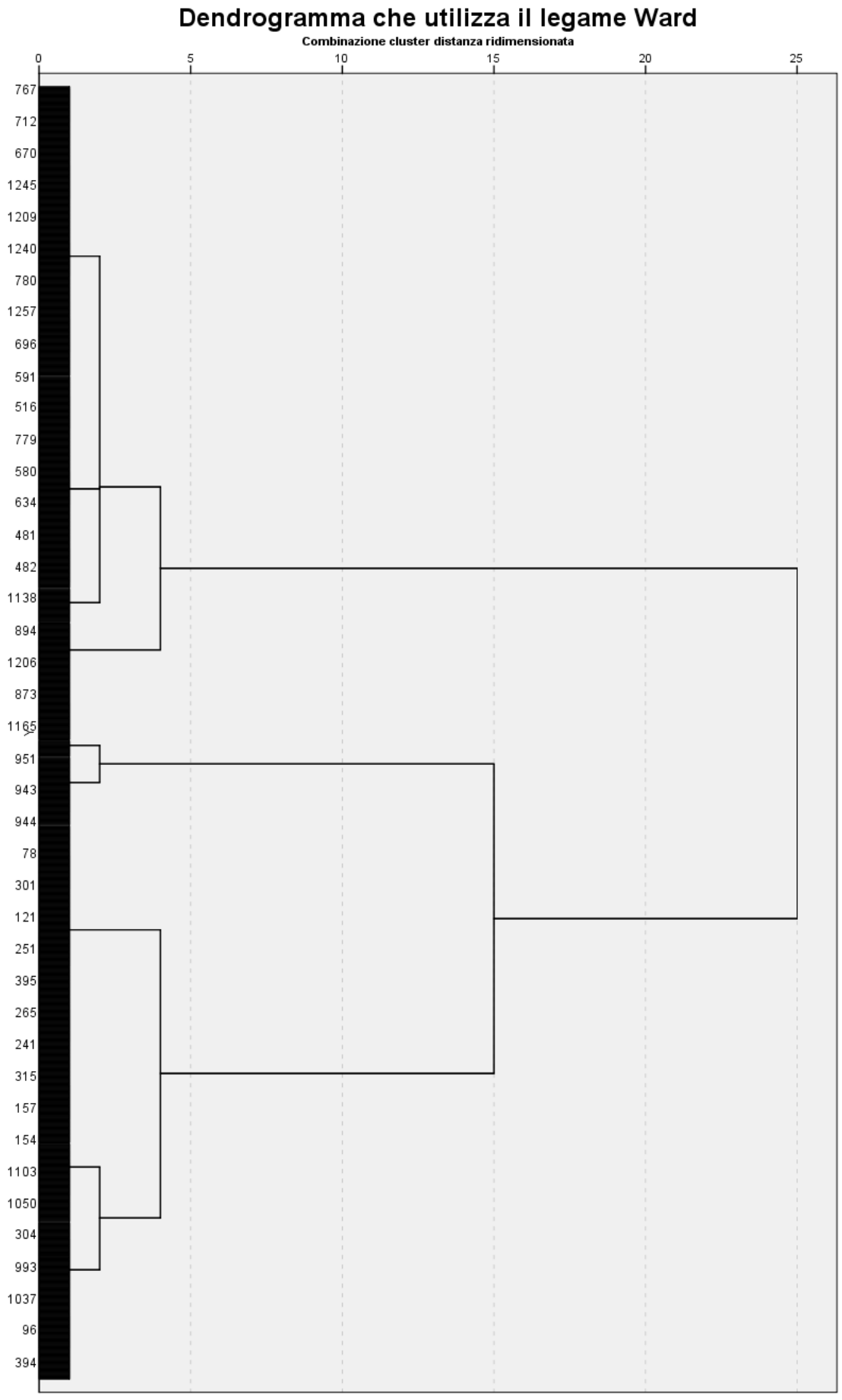
Il grafico seguente mostra l’andamento della spezzata creata con le distanze tra i cluster:

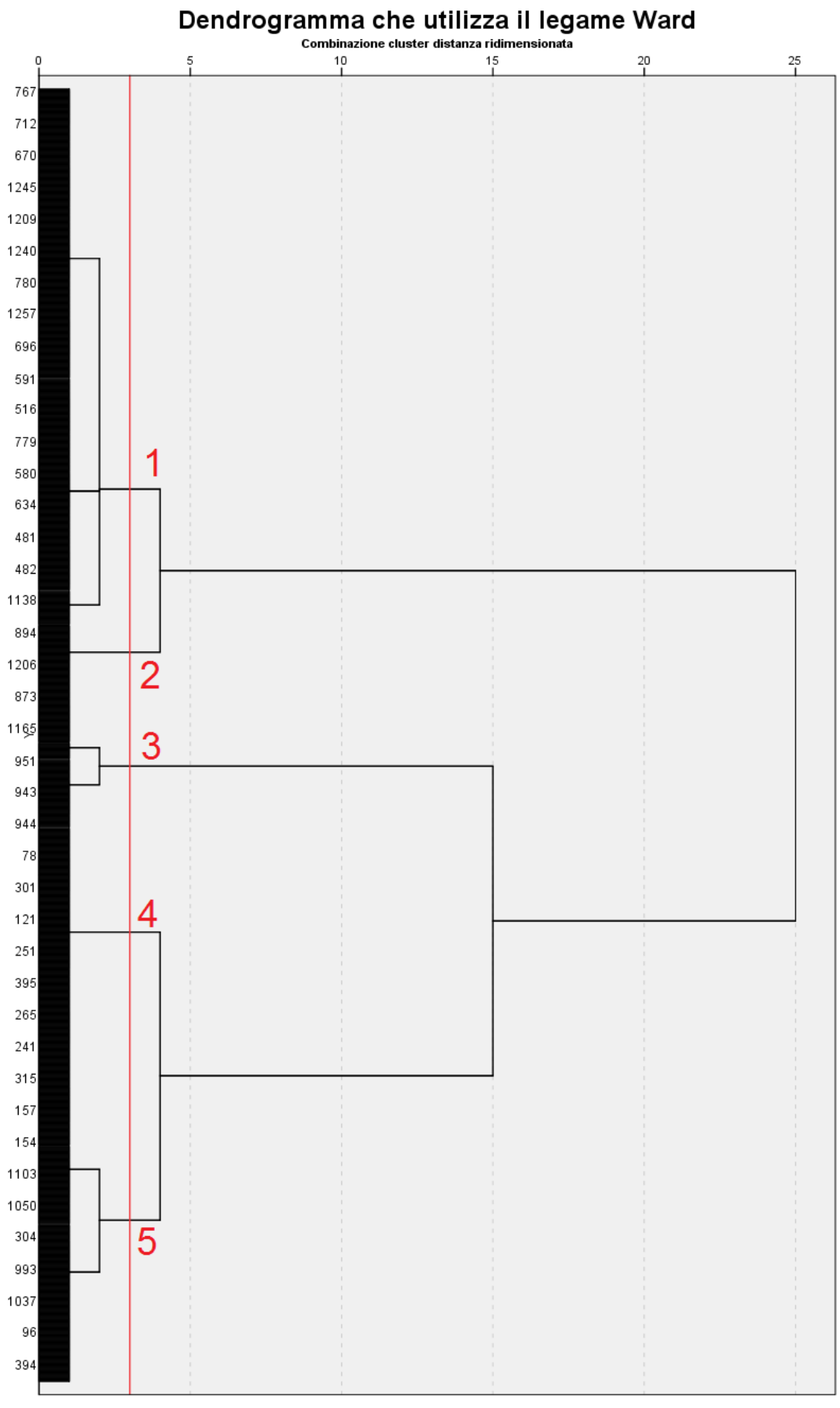


C'è una forte inclinazione nei primi raggruppamenti e un successivo appiattimento, che la porta ad essere quasi orizzontale. Noi dobbiamo cercare la soluzione a "gomito", ovvero quella che che divide meglio i due andamenti.

Nei nostri dati questo avviene tra il passo 1391 e 1392, con un aumento della distanza tra gruppi del 4.34% maggiore rispetto al passo precedente. Il numero ottimale è quindi 1396 - 1391 = 5 cluster.

Osservando il dendogramma si vedono chiaramente i nostri 5 cluster, a loro volta raggruppati a in 3 categorie principali. La distanza a cui avviene la terzultima aggregazione è nettamente la maggiore di tutte, ma anche la distanza tra la sestultima aggregazione è rilevante.


Dendrogramma che utilizza il legame Ward
Combinazione cluster distanza ridimensionata
0
5
10
15
20
25
767
712
670
1245
1209
1240
780
1257
696
591
516
779
580
634
481
482
1138
894
1206
873
1165
951
943
944
78
301
121
251
395
265
241
315
157
154
1103
1050
304
993
1037
96
394


Dendrogramma che utilizza il legame Ward
Combinazione cluster distanza ridimensionata
0
5
10
15
20
25
767
712
670
1245
1209
1240
780
1257
696
591
516
779
580
634
481
482
1138
894
1206
873
1165
951
943
944
78
301
121
251
395
265
241
315
157
154
1103
1050
304
993
1037
96
394
1
2
3
4
5

Abbiamo così trovato i nostri raggruppamenti.

Cluster	N° auto
1	255
2	343
3	579
4	127
5	93
Totale	**1396**

Ora l'ultimo passo è andare ad analizzare i nostri cluster ad uno ad uno.

Analisi esplorativa dei cluster

Prima di procedere ad analizzare i nostri raggruppamenti ad uno ad uno è utile cogliere le informazioni essenziali della nostra popolazione nella sua interezza.

Statistiche descrittive

	N	Minimo	Massimo	Media	Errore std.	Deviazione std.	Varianza	Asimmetria
Prezzo	1396	7750	365000	39311	906	33846	1145569302	3,631
Cilindrata	1396	0	6498	2066,62	25,658	958,644	918997,472	1,735
Alimentazione	1396	1	6	2,62	,042	1,577	2,487	,152
Emissioni	1396	0	398	147,48	1,298	48,510	2353,183	1,462
VelocitàMax	1396	130	350	207,65	,905	33,799	1142,383	,781
Consumo	1396	0	17	6,12	,058	2,178	4,745	1,386

L'automobile media costa 39.311€, ha una cilindrata di 2066,62 cm^3, emette 147,48 g/Km, ha un consumo medio di 6,124 litri ogni 100 KM e raggiunge una velocità massima di 207,65 Km/h.Non ha senso parlare di media per l'alimentazione in quanto è solo una trasposizione di una variabile nominale. La moda nell'alimentazione è il Diesel, con il 46% dei casi. Le varianze sono molto elevate, cosa prevedibile dato la vasta gamma di automobili che sono sul mercato.Informazioni utili ci vengono date anche dai percentili.

Percentili

	Percentili						
	5	10	25	50	75	90	95
Media ponderata	12867,60	15270,00	21000,00	29998,00	44345,75	69439,30	102948,55
	1120,00	1199,00	1497,00	1968,00	2182,75	2996,00	4371,05
	1,00	1,00	1,00	3,00	4,00	4,00	4,00
	98,00	105,00	119,00	138,00	160,00	209,00	242,00
	162,00	170,00	182,00	202,00	230,00	250,00	250,00
	3,800	4,100	4,700	5,700	6,900	8,800	10,300

Scopriamo da questa tabella ad esempio che solo il 5% dei veicoli ha prezzo superiore a 102.948€ oppure che il 75% dei modelli ha un consumo di 6.9 litri ogni 100/km (il che è molto poco sopra la media totale). Ora che conosciamo le informazioni generali possiamo confrontarle con quelle specifiche dei cluster.

Cluster 1

Statistiche descrittive

	N	Minimo	Massimo	Media	Errore std.	Deviazione std.	Varianza	Asimmetria
prezzo	255	17400	99898	42833,76	980,991	15665,165	245397409,263	1,330
cilindrata	255	1364	3778	2258,99	39,636	632,940	400613,071	,722
alimentazione	255	1	2	1,09	,018	,287	,082	2,878
emissioni	255	111	282	170,52	2,050	32,734	1071,511	,828
velocità	255	136	266	231,46	1,232	19,666	386,746	-1,055
consumi	255	4,8	12,1	7,507	,0874	1,3952	1,947	,745

Percentili

	Percentili						
	5	10	25	50	75	90	95
Media ponderata	22898,00	26964,00	31450,00	40222,00	50200,00	60433,20	77375,40
	1551,00	1596,00	1798,00	1995,00	2979,00	2996,00	3498,00
	1,00	1,00	1,00	1,00	1,00	1,00	2,00
	128,60	134,00	147,00	164,00	189,00	220,00	238,20
	195,00	201,20	220,00	235,00	250,00	250,00	250,00
	5,780	5,900	6,400	7,400	8,200	9,600	10,340

Il cluster 1 contiene 255 veicoli, è il terzo cluster più grande che è stato trovato e, come si vedrà, il meno specifico.

L'auto media del primo cluster costa circa 42.000€, è di cilindrata 2.200 cm^3, è alimentata a benzina, emette 170 g/Km, consuma 7 litri di carburante ogni 100 Km ed ha una velocità massima di 230 Km/h. Supera quindi di poco la media generale in tutti i parametri. Questa differenza è probabilmente dovuta all'alimentazione, in quando le auto a benzina hanno prestazioni più elevate rispetto agli altri tipi di alimentazione.

Si parla di auto medie a benzina, come la Mercedes Classe C 180 (36.488€, 1595 cm^3, Benzina, emissioni 136 g/Km, velocità massima 225 Km/h e 5,8 litri ogni 100 Km) oppure l'Audi A3 1.8 TFSI (27.080€, 1798 cm^3, Benzina, emissioni 135 g/Km, velocità massima 232 Km/h e 5,8 litri ogni 100 Km).

Rientrano nel cluster anche le berline sportive, che come parametri sono molto più simili alle auto medie che alle berline. Alcuni esempi sono la Abarth 500 (17.946€, 1368 cm^3, GPL, emissioni 155

g/Km, velocità massima 205 Km/h e 6,5 litri ogni 100 Km) o la Volkswagen Golf 1.4 TSI (20.900€, 1390 cm^3, Benzina, emissioni 123 g/Km, velocità massima 203 Km/h e 5,3 litri ogni 100 Km).

Il problema principale è che nel cluster finiscono anche alcuni veicoli che con l'automobile media del gruppo hanno poco in comune, ovvero le auto grandi a benzina. Ci sono una sessantina di modelli che superano i 50.000€ di prezzo, partendo dall'Audi TTS Coupé 2.0 272 (50.520€, 1984 cm3, Benzina, emissioni 184 g/Km, velocità massima 250 Km/h e 7,9 litri ogni 100 Km) fino ad arrivare alla Mercedes SL 350 (99.898€, 3498 cm3, Benzina, emissioni 159 g/Km, velocità massima 250 Km/h e 6,8 litri ogni 100 Km).

La presenza di queste auto nel raggrupppamento è dovuta all'alimentazione. Esiste infatti un Cluster (il quarto) per le auto di grandi dimensioni, queste però sono caratterizzate dall'alimentazione Diesel. Avendo il software a disposizione solo due componenti principali su cui svolgere l'analisi, è comprensibile che trovi più distanza tra auto grandi di diversa alimentazione che tra auto grandi e auto medie a benzina. Il fatto che sia giustificato però non lo rende meno sbagliato, in quanto questa categoria meriterebbe un cluster proprio.

Le case produttrici più frequenti sono BMW (59 modelli), Audi (57 modelli) e Mercedes (40 modelli).

Cluster 2

Statistiche descrittive

	N	Minimo	Massimo	Media	Errore std.	Deviazione std.	Varianza	Asimmetria
prezzo	343	7750	41944	18602,46	309,387	5729,923	32832021,302	,560
cilindrata	343	875	1998	1348,40	12,272	227,277	51655,012	-,007
alimentazione	343	1	3	1,11	,017	,320	,102	2,805
emissioni	343	85	189	131,22	1,006	18,635	347,247	,120
velocità	343	140	229	182,38	,932	17,270	298,243	,221
consumi	343	4,0	9,8	5,913	,0570	1,0558	1,115	,932

Percentili

	Percentili						
	5	10	25	50	75	90	95
Media ponderata	10160,00	11554,00	14000,00	18456,00	22182,00	25418,00	29181,20
	996,40	999,00	1197,00	1368,00	1591,00	1598,00	1598,00
	1,00	1,00	1,00	1,00	1,00	2,00	2,00
	99,00	106,40	119,00	132,00	144,00	156,20	161,60
	155,00	160,00	170,00	181,00	195,00	204,60	214,80
	4,300	4,700	5,200	5,900	6,400	7,200	8,200

Il cluster 2 contiene 343 veicoli ed è, per popolazione, il secondo raggruppamento più grande uscito dall'analisi.

L'auto media del secondo cluster costa circa 18.000€, è di bassa cilindrata (1.300 cm^3) ed è alimentata a benzina. Emette 131 g/Km, ha una velocità massima di 182 Km/h e consuma circa 5,9 litri ogni 100 km. Rispetto alla media generale le automobili in questo cluster sono molto più piccole ed economiche, ma hanno emissioni, consumi e velocità massima solo leggermente inferiori a quelli generali. Questo sempre per l'influenza dell'alimentazione a benzina, che garantisce prestazioni migliori in quanto a velocità ma è meno efficiente sui consumi e le emissioni.

Dei 5 cluster trovati è quello più chiaramente identificabile. Si parla di piccole berline, le macchine che si vedono più spesso sulle strade: Citroen C3, Ford Fiesta, Honda Civic, Opel Corsa, Renaul Clio, Toyota Yaris etc etc. È il cluster con la maggior presenza di automobili italiane, in quanto quasi tutte le auto Fiat (500, Panda, Punto, Sedici), Lancia e Alfa rientrano in questo cluster.

Rientrano in questo Cluster anche le City Car: Suzuki Alto, Toyota Aygo, Seat Mii, Smart ForTwo. Tutte queste auto hanno emissioni tra 99 e 120 g/Km, velocità massima tra i 145 e i 170 Km/h e consumano tra i 4 e 5 litri ogni 100 km.

Essendo un raggruppamento ben noto anche prima della nostra analisi non è di particolare interesse andare ad analizzare queste macchine. Può essere utile però andare a vedere i limiti che l'analisi solta pone a questa categoria, ovvero: fino a che punto un auto viene considerata una piccola berlina?

La discriminante principale in questa categoria sembra essere la cilindrata. Delle 343 macchine nella categoria, solo 10 superano i 1.600 cm^3 di cilindrata. Tutti quei modelli sono caratterizzati da bassi consumi e velocità media molto bassa (inferiore ai 200 Km/h).

L'auto più costosa del gruppo è la Volvo S80 T4 ed è, come prevedibile, di cilindrata molto bassa (solo 1.595 cm^3). L'auto più inquinante di tutti invece è un piccolo SUV: la Suzuki Grand Vitara, che con la sua cilindrata di 1586 cm^3 ed un prezzo di 23.800€ è molto più simile alle berline che ai SUV veri e propri. L'auto che consuma di più invece è la Volkswagen Golf Plus, che è una berlina sportiva ed è perfetta per questa categoria.

Le case automoblistiche più frequenti nel gruppo sono Opel (27 veicoli), Seat (22 veicoli) e Skoda (22 veicoli).

Cluster 3

Statistiche descrittive

	N	Minimo	Massimo	Media	Errore std.	Deviazione std.	Varianza	Asimmetria
prezzo	579	9750	66500	29507,46	399,430	9611,264	92376386,782	,543
cilindrata	579	0	2494	1777,65	16,548	398,174	158542,631	-1,885
alimentazione	579	3	6	4,08	,019	,462	,213	2,941
emissioni	579	0	194	121,5(	1,163	27,991	783,482	-1,331
velocità	579	130	250	196,04	,886	21,324	454,733	-,139
consumi	579	,0	7,3	4,641	,0436	1,0503	1,103	-1,424

Percentili

	Percentili						
	5	10	25	50	75	90	95
Media ponderata	15050,00	17800,00	22540,00	28500,00	36050,00	42960,00	46630,00
	1248,00	1396,00	1560,00	1968,00	1995,00	2143,00	2231,00
	4,00	4,00	4,00	4,00	4,00	4,00	5,00
	88,00	96,00	109,00	119,00	137,00	155,00	166,00
	163,00	169,00	180,00	195,00	210,00	225,00	230,00
	3,400	3,800	4,200	4,600	5,200	5,900	6,300

Con i suoi 579 veicoli il cluster 3 è nettamente quello più popolato tra quelli trovati.

L'auto media del terzo cluster costa circa 29.000€, ha una cilindrata di 1.777 cm^3, è alimentata a Diesel, emette 121 g/Km, raggiunge una velocità massima di poco inferiore ai 200 Km/h e consuma solo 4,6 litri di carburante ogni 100 km. Rispetto alle media generale quindi costa circa 10.000€ meno, ha una cilindrata inferiore ed è significativamente più ecologica: ha consumi ed emissioni molto bassi, perfino minori delle piccole berline.

Nonostante sia il cluster più numeroso offre comunque una buona classificazione. La maggior parte delle auto in questa categoria sono macchine di medie dimensioni ma ecologiche. Alcuni esempi sono la Ford Focus 2.0 TDCI (27.250€, 1997 cm^3, Diesel, emissioni 124 g/Km, velocità massima 218 e 4,9 litri di carburante ogni 100 Km) oppure Peugeot 508 2.0 140 CV (27.850€, 1997 cm^3, Diesel, emissioni 114 g/Km, velocità massima 210 e 4,4 litri di carburante ogni 100 Km).

Sono presenti molte auto familiari e station wagon, come la Renault Scenìc, Ford Mondeo, Skoda Octavia e Volkswagen Passat.

Le variabili più importanti in questa categoria sono chiaramente l'alimentazione, le emissioni e i consumi. Questo fa sì che entrino in questa categoria anche:

- auto di cilindrata molto piccola, alimentate a diesel o metano, con emissioni e consumi più vicini alle auto di questo cluster che alle normali berline. Ad esempio la Volkswagen Up 1.0 (12.400€, 999 cm^3, Metano, emissioni 79 g/Km, velocità massima 164 e 2,9 metri cubi ogni 100 Km).

- auto elettriche, che non inquinano in quanto non emettono CO_2. Un esempio è la BMW i3 (36.599€, 0 cm^3, Elettrica, emissioni 0 g/Km, velocità massima 150 e 0 consumi).

- auto costose, alimentate a diesel oppure ibride, ma ecologiche e dalle basse prestazioni. Ad esempio la Lexus GS Hybrid (61.000€, 2494 cm3, Ibrida, emissioni 115 g/Km, velocità massima 190 e 4,9 litri ogni 100 Km), oppure la Mercedes E 250 CDI (53.316€, 2143 cm3, Diesel, emissioni 136 g/Km, velocità massima 232 e 5,2 litri ogni 100 Km).

In un analisi basata esclusivamente sulle prestazioni queste auto non andrebbero nello stesso cluster, ma avendo basato gran parte dell'analisi su queste variabili (alimentazione, emissioni e consumi) ha perfettamente senso un cluster di questo tipo. Un cluster formato da macchine ecologiche di piccola e media taglia.

Le case produttrici più frequenti nella categoria sono BMW (48 modelli), Opel (41 modelli), Audi (39 modelli).

Cluster 4

Statistiche descrittive

	N	Minimo	Massimo	Media	Errore std.	Deviazione std.	Varianza	Asimmetria
prezzo	127	30240	126150	63928,59	1743,699	19650,492	386141836,307	,999
cilindrata	127	1984	4969	3066,02	41,434	466,934	218027,420	1,710
alimentazione	127	4	6	4,31	,065	,731	,535	1,903
emissioni	127	129	250	176,10	2,864	32,273	1041,521	,496
velocità	127	172	270	230,91	2,172	24,472	598,864	-1,092
consumi	127	4,9	9,9	6,802	,1081	1,2183	1,484	,476

Percentili

	Percentili						
	5	10	25	50	75	90	95
Media ponderata	38427,80	44130,80	50720,00	58080,00	74850,00	93536,80	105830,40
	2400,00	2776,00	2967,00	2987,00	2993,00	3498,00	4134,00
	4,00	4,00	4,00	4,00	4,00	6,00	6,00
	134,00	138,80	149,00	166,00	199,00	224,00	232,80
	177,00	186,00	220,00	240,00	250,00	250,00	250,00
	5,100	5,300	5,900	6,400	7,900	8,520	9,000

Il cluster 4 contiene 127 veicoli.

L'auto media del gruppo ha un prezzo di circa 60.000€, ha un motore cilindrata 3.000 cm^3, è alimentata a Diesel, raggiunge una velocità massima di 230 km/h emettendo 176 g/Km e consumando 6,8 litri di caruburante ogni 100 km. È quindi un auto molto più costosa della media generale e di cilindrata maggiore, che inquina di più ma ha prestazioni paragonabili.

È un cluster molto specifico e contiene sostanzialmente due tipi di auto:

- auto dirigenziali: costose, eleganti e poco sportive. Alcuni esempi sono l'Audi A5 3.0 (46.690€, 2967 cm^3, Diesel, emissioni 129 g/Km, velocità max 235 e 4,9 litri ogni 100 km) oppure la BMW 325 Eletta (51.522€, 2993 cm^3, Diesel, emissioni 160 g/Km, velocità max 238 e 6,1 litri ogni 100 km).

- SUV: sono tutti in questa categoria, da quelli più economici come la Jeep Cheroke (30.240€) a quelli più costosi, come la BMW X6 (95.483€) e la Porsche Cayenne (84.658€). La

selezione in questo caso è veramente molto buona in quanto tutti i SUV hanno più o meno le stesse caratteristiche a prescindere dal prezzo (il che spiega anche gran parte della varianza di questa variabile).

Questi due tipi di auto possono sembrare molto diverse alla vista ma sono alquanto simili nei parametri. Inoltre identificano più o meno lo stesso tipo di consumatore (che potremmo definire "alto borghese") quindi è accettabile che stiano insieme.

Rientrano in questa categoria anche modelli i modelli di lusso che, essendo a motore ibrido, hanno parametri più vicini alle auto dirigenziali che alle auto del cluster 5. In particolare hanno emissioni e consumi molto più bassi di quella categoria. Alcuni esempi sono la Mercedes S 400 (94.000€, 3.498 cm^3, Ibrida, emissioni 147 g/Km, velocità massima 250, 6,3 litri ogni 100 Km) e la Peugeot Panamera (117.000€, 2995 cm^3, ibrida, emissioni 167 g/Km, velocità massima 270, 7,1 litri ogni 100 Km).

Le case più frequenti nella categoria sono Audi (32 veicoli), BMW (23 veicoli) e Mercedes (17 veicoli).

Cluster 5

Statistiche descrittive

	N	Minimo	Massimo	Media	Errore std.	Deviazione std.	Varianza	Asimmetria
prezzo	93	40920	365000	133816,99	6021,209	58066,435	3371710834,446	1,525
cilindrata	93	2979	6498	4650,24	89,059	858,858	737637,161	,116
alimentazione	93	1	4	1,03	,032	,311	,097	9,644
emissioni	93	195	398	267,60	5,391	51,991	2703,090	,443
velocità	93	175	350	276,56	3,503	33,784	1141,358	,100
consumi	93	8,4	17,2	11,449	,2304	2,2221	4,938	,469

Percentili

	Percentili						
	5	10	25	50	75	90	95
Media ponderata	71808,70	75938,80	94430,00	118635,00	156805,00	212668,20	252541,80
	2997,00	3436,00	3993,00	4663,00	5204,00	6089,20	6224,20
	1,00	1,00	1,00	1,00	1,00	1,00	1,00
	199,00	206,80	224,00	262,00	304,50	349,20	355,80
	242,50	250,00	250,00	278,00	304,50	322,40	328,00
	8,600	8,840	9,500	11,200	13,100	14,780	15,430

Il Cluster 5 contiene 93 veicoli ed è il più piccolo di quelli trovati

L'auto media del quinto cluster costa ben 133.000€, ha una cilindrata di 4.650 cm^3, è alimentata a Benzina (anche se sono presenti anche veicoli a Diesel). Emissioni, velocità massima e consumi sono molto elevati (267 g/Km, 274 km/h, 11,5 litri ogni 100 km). Sono veicoli quindi che superano la media generale in tutte le categorie.

Anche in questo cluster contiene due tipi di veicoli:

- auto grandi con prestazioni sportive, come l'Infiniti QX70 5.0 (72.850€, 5026 cm^3, Benzina, emissioni 307 g/Km, velocità massima 250 Km/h, 13,1 litri ogni 100 Km) oppure la Chevrolet Camaro V8 (40920€, 6162 cm^3, Benzina, 329 g/Km, velocità massima 250 Km/h, 14 litri ogni 100 Km).

- auto di lusso. Ferrari, Jaguar, Lamborghini, Mercedes: quasi tutti i modelli di questi case automobilistiche sono in questa categoria.

Ovviamente la seconda cateoria è compresa inevitabilmente nella prima, in quanto tutte le auto di lusso hanno prestazioni uguali o maggiori delle auto sportive.

Come nel cluster precedente, anche qui ha senso mettere queste due categorie di auto insieme, in quanto, tolto il prezzo, hanno parametri molto simili. La differenza di prezzo è però molto rilevante perché identifica pubblici completamente diversi: chi compra una macchina sportiva da 60.000€ probabilmente vorrebbe avere una Ferrari, ma difficilmente può permettersela.

Sono presenti alcuni errori di classificazione all'interno di questo cluster. Per esempio la Mercedes G 350 rientra in questo cluster, ma è un SUV e in quanto tale dovebbe stare nel Cluster 4. È facile capire perché non sia stato messo insieme agli altri SUV, in quanto inquina e consuma quasi il doppio degli altri fuoristrada, ma ha una cilindrata (2987 cm^3) e una velocità massima (175 Km/h) veramente troppo bassi per stare nel gruppo delle auto dalle grandi prestazioni.

Per il resto è un gruppo conosciuto e ben riconoscibile. I parametri più importanti in questo cluster sono chiaramente quelli legati alle prestazioni. Il fatto che quasi tutti i veicoli siano a benzina è una semplice conseguenza del fatto che la benzina è il tipo di alimentazione più indicato per poter raggiungere quelle prestazioni.

Le case più frequenti nella categoria sono BMW (48 veicoli), Mercedes (45 veicoli) e Audi (29 veicoli).

Conclusioni

L'analisi dei dati tramite il software SPSS è stata svolta senza problemi e la procedura che si è scelto si è rivelata adatta per classificare il mercato delle auto.

Per quanto riguarda la bontà della soluzione ottenuta, bisogna prima di tutto considerare che riuscire a classificare quasi 1400 casi è molto complicato: anche la migliore classificazione possibile avrebbe sicuramente dei casi anomali all'interno dei cluster.

Considerando l'elevato numero di dati che avevamo a disposizione possiamo ritenerci abbastanza soddisfatti dai cluster ottenuti. Ogni gruppo infatti rappresenta una macro-categoria facilmente identificabile:

- Il cluster 1 è caratterizzato dalle auto medie alimentate a benzina.
- Il cluster 2 è caratterizzato dalle berline piccole.
- Il cluster 3 è caratterizzato dalle auto medie, alimentate a Diesel e poco inquinanti.
- Il Cluster 4 è caratterizzato dalle auto grandi alimentate a Diesel.
- Il Cluster 5 è caratterizzato dalle auto di lusso a benzina.

Cinque categorie però sono troppo poche per rappresentare il mercato automobilistico nella sua interezza. Si sarebbe dovuto probabilmente andare più a fondo con l'analisi, o scegliendo una soluzione con più cluster durante l'elaborazione oppure effettuando una nuova analisi su alcuni gruppi ottenuti.

Il cluster 1 e il cluster 5 contengono infatti due categorie di auto che la nostra analisi così com'è non tiene in conto: le auto grandi alimentate a Benzina e le auto grandi dalle alte prestazioni. Due categorie importanti in quanto identificano consumatori diversi rispetto alla macchina media del gruppo.

Per il resto possiamo ritenerci soddisfatti della soluzione ottenuta. Aver svolto l'analisi sia su variabili classiche (come il prezzo e la cilindrata) sia su variabili variabili particolari (come l'alimentazione e le emissioni) ha permesso di classificare molti modelli nel gruppo più corretto per loro.

Certo i cinque cluster trovati sono raggruppamenti tutto sommato già noti nel mondo reale, ma una soluzione più estrema sarebbe risultata poco realistica. In verità, essere arrivati a questa soluzione, utilizzando variabili raramente usate in questo tipo di analisi, in qualche modo ci conferma la bontà del modello.

Bibliografia

Aragona B., *Tecniche di analisi multivariata: alcune applicazioni con SPSS*, Napoli, Liguori, 2013.

Everitt B. - Landau S. - Morven L., *Cluster Analysis 5th ed.,* Hoboken, Wiley, 2011.

Fabbris L., *Statistica multivariata. Analisi esplorativa dei dati*, Milano, McGraw-Hill, 1997.

Molteni L., *L'analisi multivariata nelle ricerche di marketing. Applicazioni alla segmentazione della domanda e al mapping multidimensionale*, Milano, Egea, 1993.

Morgan G. - Leech N. - Gloeckner G. - Barrett K., *IBM SPSS for introductory statistic: use and interpretation*, New York, Routledge, 2013.

Sitografia

Analisi dei dati con SPSS, di Barbanelli Claudio, in
http://www.lededizioni.com/lededizioniallegati/barbaranellispss_1.pdf

Cluster gerarchica, in
http://host.uniroma3.it/facolta/economia/db/materiali/insegnamenti/586_5037.pdf

I segmenti vettura nel settore auto, in
http://marketingandstyle.blogspot.it/2012/03/i-segmenti-vettura-nel-settore-auto.html

Manuale SPSS Statistics Base 20, in
ftp://public.dhe.ibm.com/software/analytics/spss/documentation/statistics/20.0/it/client/Manuals/IBM_SPSS_Statistics_Base.pdf

Segmentazione della domanda e scelta del target di mercato, di Andrea d'Angelo, in
http://www.disp.uniroma2.it/users/dangelo/TESTI/Fondamenti_di_Marketing/Segmentazione%20e%20scelta_target.pdf

SPSS Tutorial, in
http://www.mvsolution.com/wp-content/uploads/SPSS-Tutorial-Cluster-Analysis.pdf

Appendice A - Formule

Indice di correlazione di Pearson:

$$\rho XY = \frac{\sigma XY}{\sigma X \sigma Y}$$

Test KMO:

$$\text{KMO} = \frac{\sum_{i=1}^{k} \sum_{j=1}^{k} r_{ij}^2}{(\sum_{i=1}^{k} \sum_{j=1}^{k} r_{ij}^2 + \sum_{i=1}^{k} \sum_{j=1}^{k} a_{ij}^2)}$$

$$\text{Con } a_{\text{ij}} = (r_{\text{ij}} \bullet 1, 2, 3, \ldots, \text{k})$$

Test di sfericità di Bartlett:

$$\text{T} = \frac{(N-k)\ln(S_p^2) - \sum_{i=1}^{k}(ni-1)\ln(S_i^2)}{1 + \frac{1}{3(k-1)}\left(\sum_{i=1}^{k}(\frac{1}{ni-1}) - \frac{1}{N-k}\right)}$$

Distanza euclidea quadratica:

$$d\ (A,B)\ = \sum_{i=1}^{k} |X_{ai} - X_{bi}|^2$$

Metodo di raggruppamento di Ward:

$$d_{(i,j)k} = \frac{1}{ni+nj+nk}\left[(ni+nk)d_{ik}^2 + (nj+nk)\ d_{jk}^2 - n_k d_{ij}^2\right]$$

Appendice B - Tabelle

Tutte le tabelle che per ragioni di spazio non è stato possibile inserire nella tesi sono consultabili on-line.

- Dataset iniziale: *http://bit.ly/Dataset_Iniziale*
- Dataset finale, con fattori principali e cluster di appartenenza: *http://bit.ly/Dataset_Finale*
- Programma di agglomerazione completo: *http://bit.ly/Programma_Agglomerazione*
- Dendogramma esteso: *http://bit.ly/Dendogramma*