Prezzi di auto su AutoScout24

Strumenti statistici per l'analisi di dati aziendali

Mirko Gabriel Briglia, Alice Cappella, Filippo Scalabrin

Corso di Laurea Magistrale in Scienze Statistiche

A.A. 2023/2024

Indice

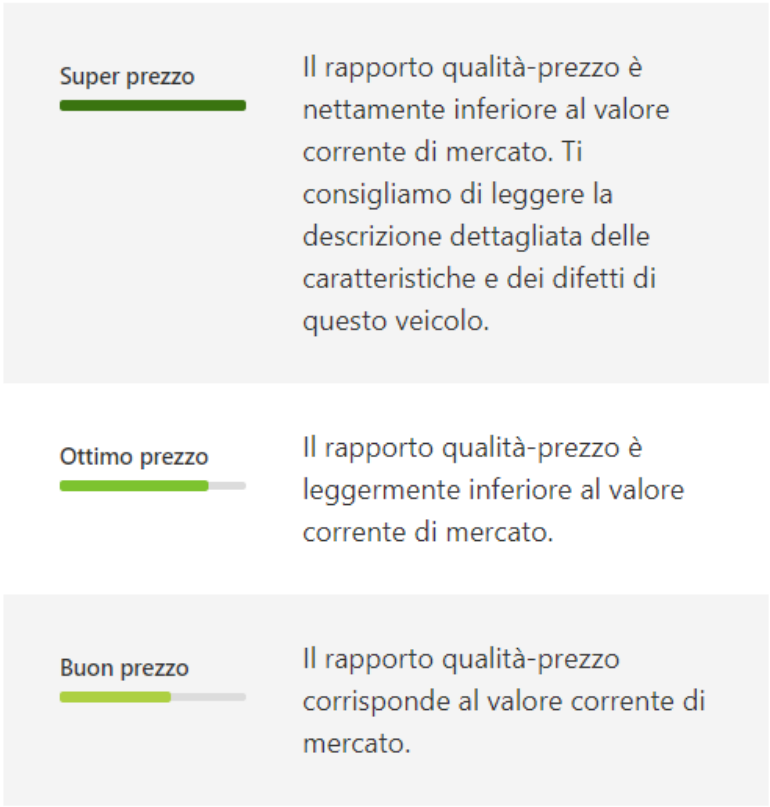
Introduzione

e domanda di business

AutoScout24

- AutoScout24 è il più grande sito web di annunci auto in Europa.
- Inserzioni pubblicabili da privati o da concessionarie.
- Per ogni veicolo possono essere presenti:
  - dati di base e cronologia interventi;
  - dati tecnici;
  - equipaggiamento (optional);
  - colori e materiali;
  - valutazione del prezzo secondo AutoScout24.

Valutazione del prezzo



- L'algoritmo di AutoScout24 calcola il prezzo di mercato per ogni veicolo inserito.
- Il prezzo di mercato viene confrontato con il prezzo finale del veicolo.
- La differenza porta all'assegnazione delle etichette di prezzo.
- Tra i criteri per il calcolo del prezzo: chilometraggio, optional, anno di immatricolazione...

Domanda di business

Le richieste di AutoScout24

- Individuare strutture di associazione tra gli optional delle auto: c'è il sospetto che gli utenti commettano errori, duplicazioni od omissioni nell'inserimento degli optional sul portale.
- Proporre uno o più modelli alternativi a quello esistente per la previsione del prezzo di una vettura, a partire dalle sue caratteristiche.
- Suggerire criteri per raggruppare le auto in base alla loro similarità.

Creazione del dataset

scraping e pulizia dei dati

Punto di partenza: i dati

#	Marca	Modello
1	Fiat	Panda
2	Dacia	Sandero
3	Lancia	Ypsilon
4	Toyota	Yaris Cross
5	Fiat	500
6	Volkswagen	T-Roc
7	Renault	Captur
8	Citroen	C3
9	Ford	Puma
10	Dacia	Duster
11	Jeep	Renegade
12	Fiat	500x
13	Renault	Clio
14	Peugeot	208
15	Peugeot	2008
16	Toyota	Yaris
17	Opel	Corsa
18	Jeep	Avenger
19	Jeep	Compass
20	Volkswagen	T-Cross
21	Nissan	Qashqai
22	MG	ZS
23	Kia	Sportage
24	Peugeot	3008
25	Alfa Romeo	Tonale

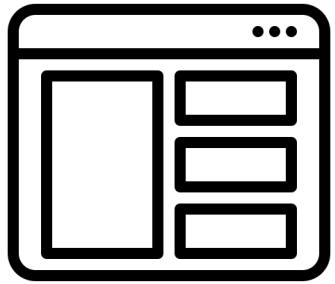
#	Marca	Modello
26	Volkswagen	Polo
27	Ford	Kuga
28	Volkswagen	Tiguan
29	Ford	Focus
30	Toyota	Aygo X
31	Audi	Q3
32	BMW	X1
33	Hyundai	i10
34	Hyundai	Tucson
35	Citroen	C3 Aircross
36	Cupra	Formentor
37	Mercedes	GLA
38	Ford	Fiesta
39	Kia	Picanto
40	Nissan	Juke
41	Audi	A3
42	Opel	Mokka
43	Suzuki	Ignis
44	Fiat	Tipo
45	Volkswagen	Golf
46	Renault	Austral
47	Audi	A1
48	Mini	Countryman
49	Volkswagen	Taigo
50	Suzuki	Vitara

Elenco delle 50 auto più vendute in Italia nel 2023 (fonte: alVolante).

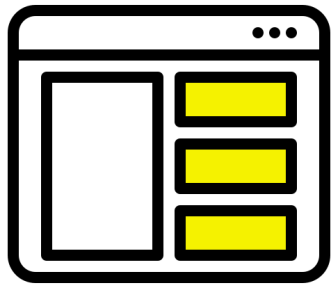
Scraping: procedura

Per ognuna delle 50 vetture più vendute in Italia...

1) Link alla pagina con i risultati di ricerca di AutoScout24.



2) Per ogni pagina di risultati di ricerca, collezione dei link ad ogni singolo annuncio.



3) Scraping dei dati della vettura corrispondente ad ogni annuncio.

Dopo lo scraping

- Lo scraping è stato fatto in due occasioni di tempo diverse (marzo e aprile).
  - È capitato di collezionare due volte la stessa unità statistica.
- Il dataset ottenuto con lo scraping ha bisogno di corpose modifiche ed è ricco di dati mancanti:
  - manipolazione delle stringhe e accorpamento di livelli per rielaborare le variabili;
  - imputazione di valori mancanti usando...
    - la logica;
    - tecniche statistiche.

Variabili eliminate

Variabili con troppi valori mancanti

Ad esempio quelle sulla manutenzione del veicolo:

- `cambio_cinghia_distribuzione`
- `ultimo_tagliando`
- `revisione`

Variabili leaker

Tutte quelle relative al pagamento rateale della vettura o agli acconti, come:

- `anticipo`
- `rata_mensile`
- `importo_lordo_del_credito`

Variabili inutili o ridondanti

Ad esempio,

- `versione_per_nazione`
- `offerta_n`
- `extra`

Variabili modificate

modello

X1 sdrive18d Futura
X1

prezzo_auto

€ 24.900,-
24900

chilometraggio

64.518 km
64518

cilindrata

1.995 cm^3
1995

anno_prod

11/2013
2013

regione

Viterbo - Vt, IT
Lazio

zona_geografica

Piemonte
Nord Ovest

potenza

51 kW (69 CV)
kW: 51 - CV: 69

per_neopatentati

Si - NA
Si - No

- A partire dalla variabile originale `potenza`, sono state ricavate le due variabili `kW` e `cv`.
- La variabile `per_neopatentati` è stata ricodificata sulla base del limite di legge di 55 kW.

peso

1.090 kg
1090

carrozzeria

SUV/Fuoristrada/Pick-up
SUV

tagliandi_certificati

Sì - NA
Si - No

tipo_di_veicolo

KM0
Nuovo

ibrida

Corrente elettrica - NA
Si - No

classe_di_emissioni

Euro 6c
Euro 6 Plus

veicolo_non_fumatori

Sì - NA
Sì - Non dichiarato

usato_garantito

17 mesi
24-36 mesi

colore_finiture_interne

Arancione
Colorate

- Dataset originale: stesso modello etichettato con tipi di carrozzeria differenti. Dopo le modifiche ogni modello ha ricevuto lo stesso tipo di carrozzeria.
- Veicoli KM0 o dimostrativi: equiparati ai nuovi.

carburante

Super Plus 98
Benzina

colore

Lilla
Colorata

metallizzata

Metallizzata - Altro
Si - No

materiale

Pelle scamosciata
Pelle

consumi

3,9 l/100 km (comb.)
3.9

emissioni

133 g/km (comb.)
133

marce

0
Automatico

- Alcune auto con cambio automatico avevano un valore di marce pari a `0` o `1`: la variabile `marce` è stata ricodificata con livelli `5`, `6 o più`, `Automatico`.

Imputazione dei valori mancanti

- Imputazione dei valori mancanti delle seguenti variabili.
  - `chilometraggio`: valore 0 per i veicoli nuovi.
  - `peso`: peso mediano del modello.
  - `tagliandi_certificati`: modalità `No` (improbabile che non venga specificato questo utile dato quando effettivamente l'auto ha fatto dei tagliandi certificati).
  - `classe_emissioni`: in base all'anno di produzione dell'auto.
  - `emissioni`: emissioni mediane del modello.
  - `consumi`: consumi medi di modello.

- `trazione`:
  1) calcolo del prezzo medio per modello e categoria di trazione;
  2) associazione in base al prezzo medio più vicino.
- `cilindri`:
  1) 0 e 1 posti come *NA*;
  2) valore più frequente.
- `marce`: numero di marce del modello col prezzo medio più simile.
- `cilindrata`:
  1) scelta variabile maggiormente correlata, `peso` (correlazione 0.667);
  2) lisciamento LOESS.
- `porte`, `posti`: modalità più frequente.
- `kW`, `cv`: valori medi di modello.

Optional

- Lo scraping degli optional di ogni auto ha restituito una stringa del tipo "`Alzacristalli elettrici*Autoradio*CD*MP3*ABS*...`".
- Ogni optional è quindi separato dall'altro da un asterisco.

Problemi

- Negli optional compaiono anche costrutti indesiderati come "`I dati di consumi ed emissioni per le auto usate si intendono riferiti al ciclo NEDC`".
- *Moltissimi* optional appartengono a *pochissimi* modelli di specifiche marche.

Soluzioni

1. Selezione dei 45 optional più rilevanti;
2. costruzione di un'indicatrice per ogni optional rilevante;
3. controllo: l'optional compare nella stringa di un'auto?

Variabili del dataset definitivo e loro descrizione

- Alle variabili sottostanti vanno aggiunte le indicatrici relative agli optional.
- In totale, il dataset conta 1591 righe e 76 colonne.

#	Variabile	Descrizione
1	`marca`	Casa costruttrice
2	`modello`	Nome del modello
3	`regione`	Regione del rivenditore
4	`zona_geog.`	Zona del rivenditore
5	`prezzo_auto`	Prezzo in Euro
6	`anno_prod`	Anno di produzione
7	`carrozzeria`	Tipo di carrozzeria
8	`tipo_di_veicolo`	Veicolo nuovo o usato
9	`trazione`	Trazione
10	`posti`	Numero di sedili
11	`porte`	Numero di porte
12	`per_neopaten.`	Per neopatentati?
13	`chilometraggio`	Numero di km percorsi
14	`kW`	Potenza in kW
15	`cv`	Potenza in CV
16	`cambio`	Tipo di trasmissione

#	Variabile	Descrizione
17	`cilindrata`	Cubatura motore
18	`cilindri`	Cilindri del motore
19	`peso_a_vuoto`	Peso del veicolo
20	`carburante`	Tipo di alimentazione
21	`classe_emissioni`	Classe ambientale
22	`colore`	Colore dell'auto
23	`metallizzata`	Colore metallizzato?
24	`materiale`	Materiale interni
25	`ibrida`	Auto ibrida?
26	`usato_garantito`	Mesi di garanzia
27	`veicolo_non_fum.`	Sì o non dichiarato?
28	`finiture`	Colore finiture
29	`tagliandi_certif.`	Prova dei tagliandi?
30	`emissioni`	CO_2 emessa in g/km
31	`consumi`	Consumi (test WLTP)
...	tutti gli optional	

- Sono stati scartati gli optional:
  - troppo diffusi (la cui presenza è data per scontata nelle auto più costose);
  - troppo poco diffusi (peculiarità esclusiva di una certa marca o modello).
- Di seguito, gli optional per i quali sono state costruite le variabili indicatrici.

#	Optional
32	autoradio
33	isofix
34	park.distance.control
35	cruise.control
36	fendinebbia
37	bluetooth
38	volante.in.pelle
39	climatizzatore
40	sensori.parcheggio
41	volante.multifunzione
42	usb
43	computer.di.bordo
44	c.pressione.gomme
45	sedile.post.sdoppiato
46	cerchi.in.lega

#	Optional
47	c.automatico.trazione
48	immobilizer
49	airbag.testa
50	bracciolo
51	sensore.luminosita
52	start.stop
53	navigatore
54	autoradio.digitale
55	frenata.emergenza
56	touch.screen
57	chiusura.c.telecom.
58	sensore.pioggia
59	clima.automatico
60	mp3
61	fari.led

#	Optional
62	vivavoce
63	controllo.el.corsia
64	android.auto
65	apple.car.play
66	hill.holder
67	telecamera.posteriore
68	cruise.control
69	vetri.oscurati
70	sound.system
71	luci.diurne.led
72	luci.diurne
73	ricon.segnali
74	sens.parcheggio.ant.
75	kit.antipanne
76	antifurto

Analisi esplorativa

...e consigli per l'acquisto!

Che macchina comprare?

- ...no, questa non è presente nel dataset.

Che macchina comprare?

Le più costose

1	Alfa Romeo Tonale	€ 32.907
2	Renault Austral	€ 28.275
3	Cupra Formentor	€ 26.957

Le più economiche

1	Fiat Panda	€ 8.320
2	Fiat 500	€ 8.729
3	Lancia Ypsilon	€ 8.841

Dotazione meno ricca

1	Fiat Panda	7
2	Kia Picanto	11
3	Lancia Ypsilon	12

Dotazione più ricca

1	MG ZS	28,5
2	Peugeot 3008	27
3	Ford Puma	27

Le più assetate

1	Kia Sportage	6,04 l/100km
2	MG ZS	5,82 l/100km
3	Volkswagen Tiguan	5,75 l/100km

Le salva-portafoglio

1	Toyota Yaris	3,55 l/100km
2	Toyota Aygo-X	4,00 l/100km
3	Peugeot 3008	4,15 l/100km

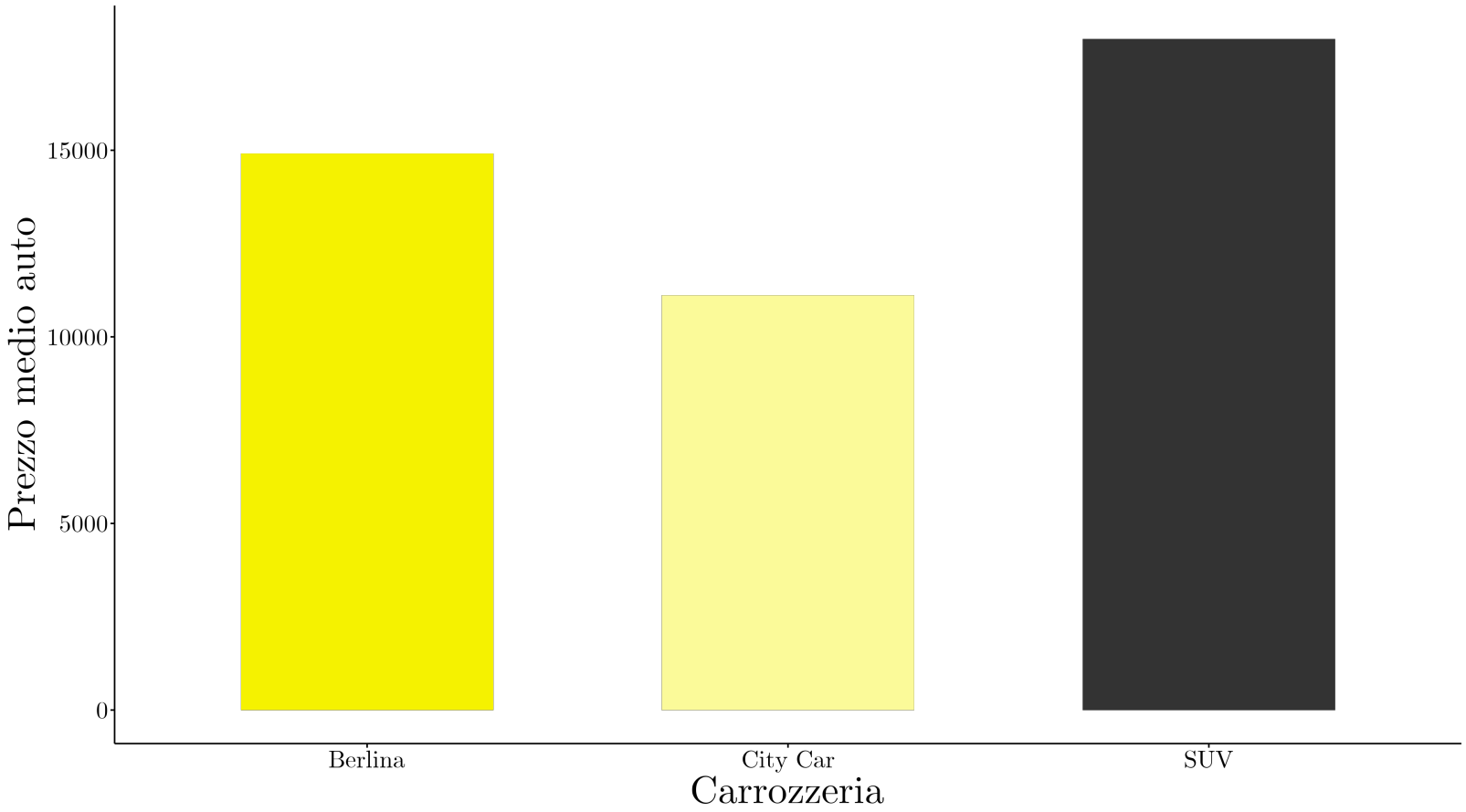
Le meno potenti

1	Kia Picanto	66 CV
2	Fiat Panda	67 CV
3	Hyundai i10	69 CV

Le più potenti

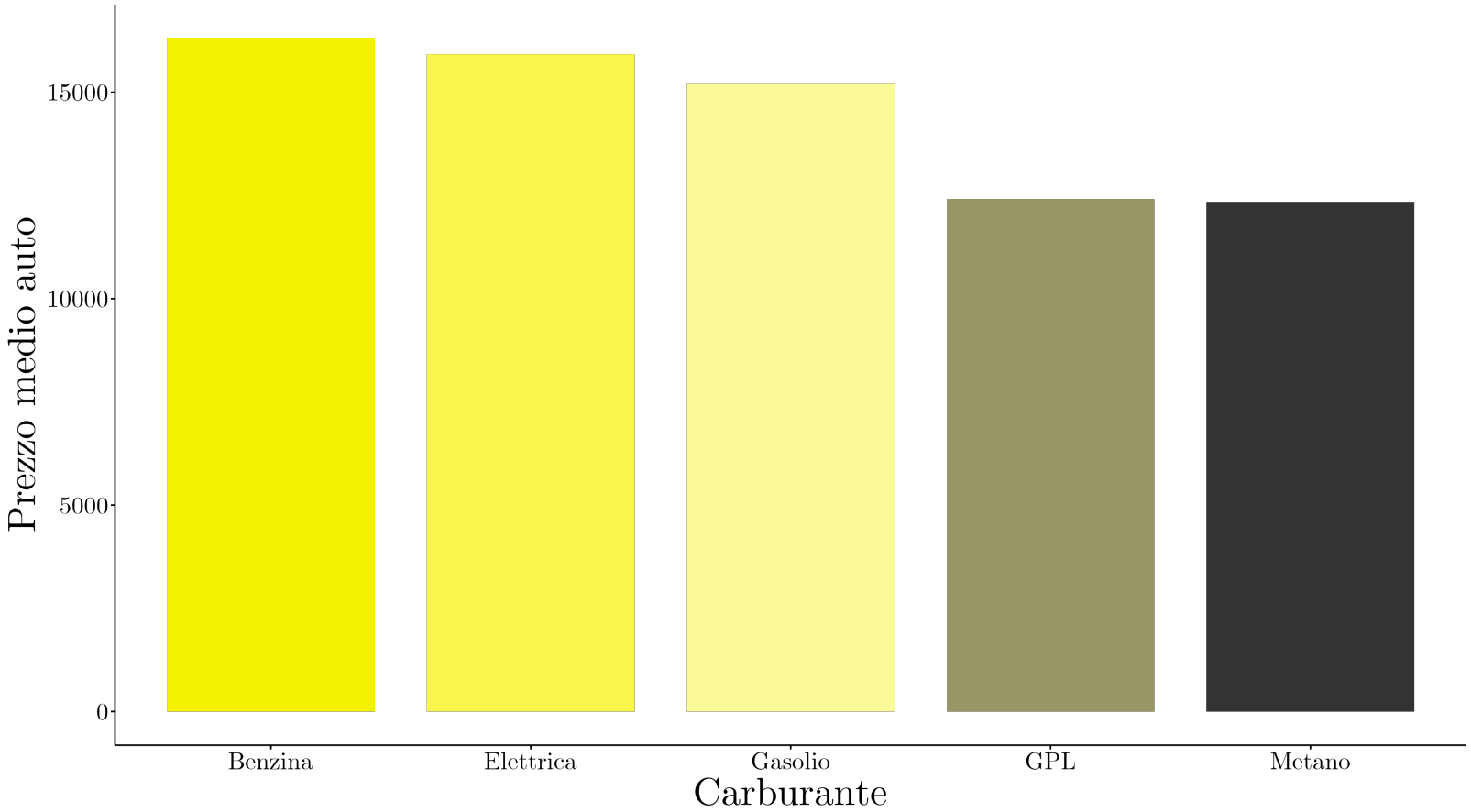
1	Mercedes GLA	160 CV
2	BMW X1	158 CV
3	Audi Q3	155 CV

Prezzo medio per tipologia di auto



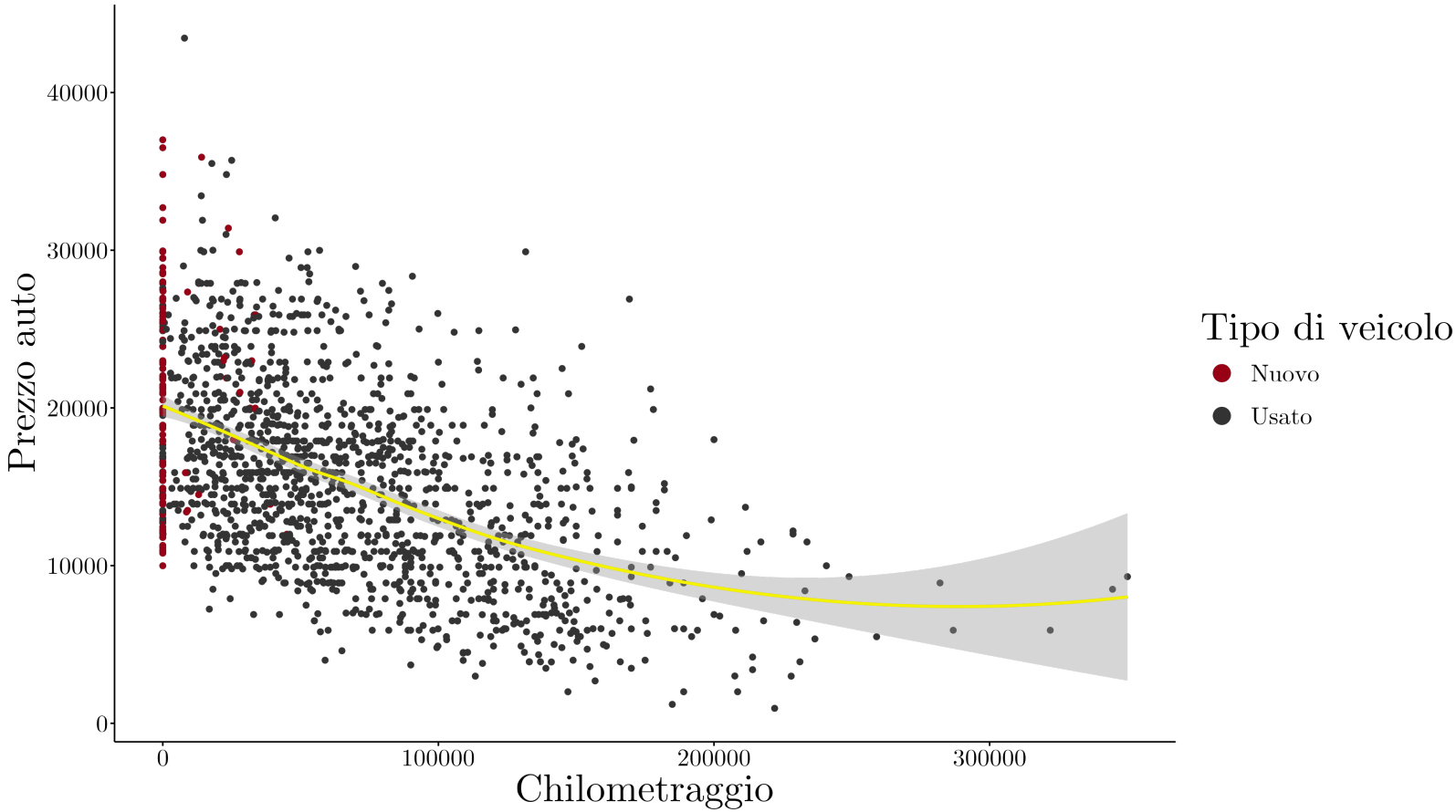
- Tra il prezzo medio dei SUV e quello delle city car c'è una differenza di quasi 7000 €.

Prezzo medio per tipologia di carburante



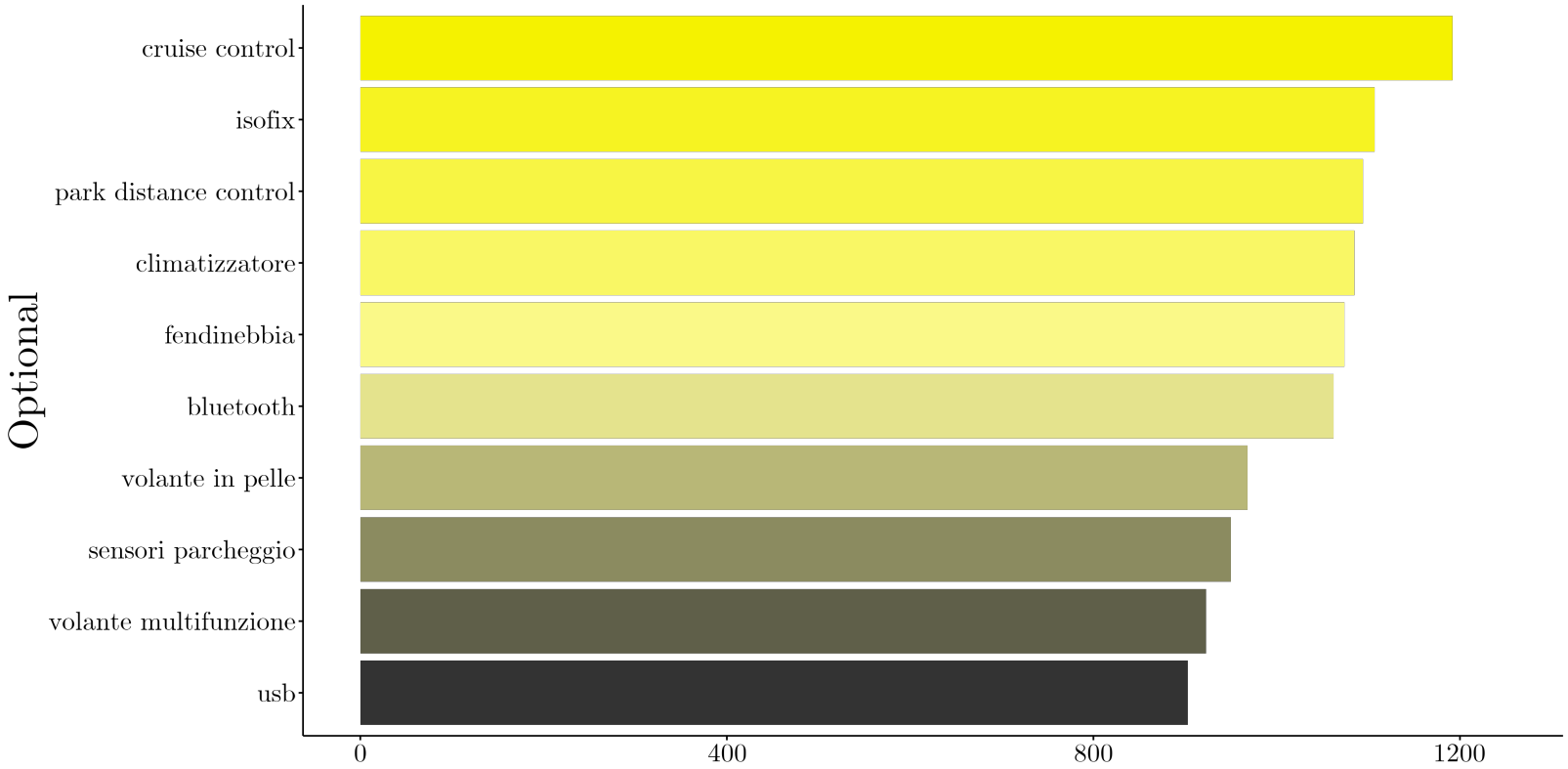
- Le elettriche (usate) costano meno delle auto a benzina (usate).

Relazione tra prezzo e chilometraggio



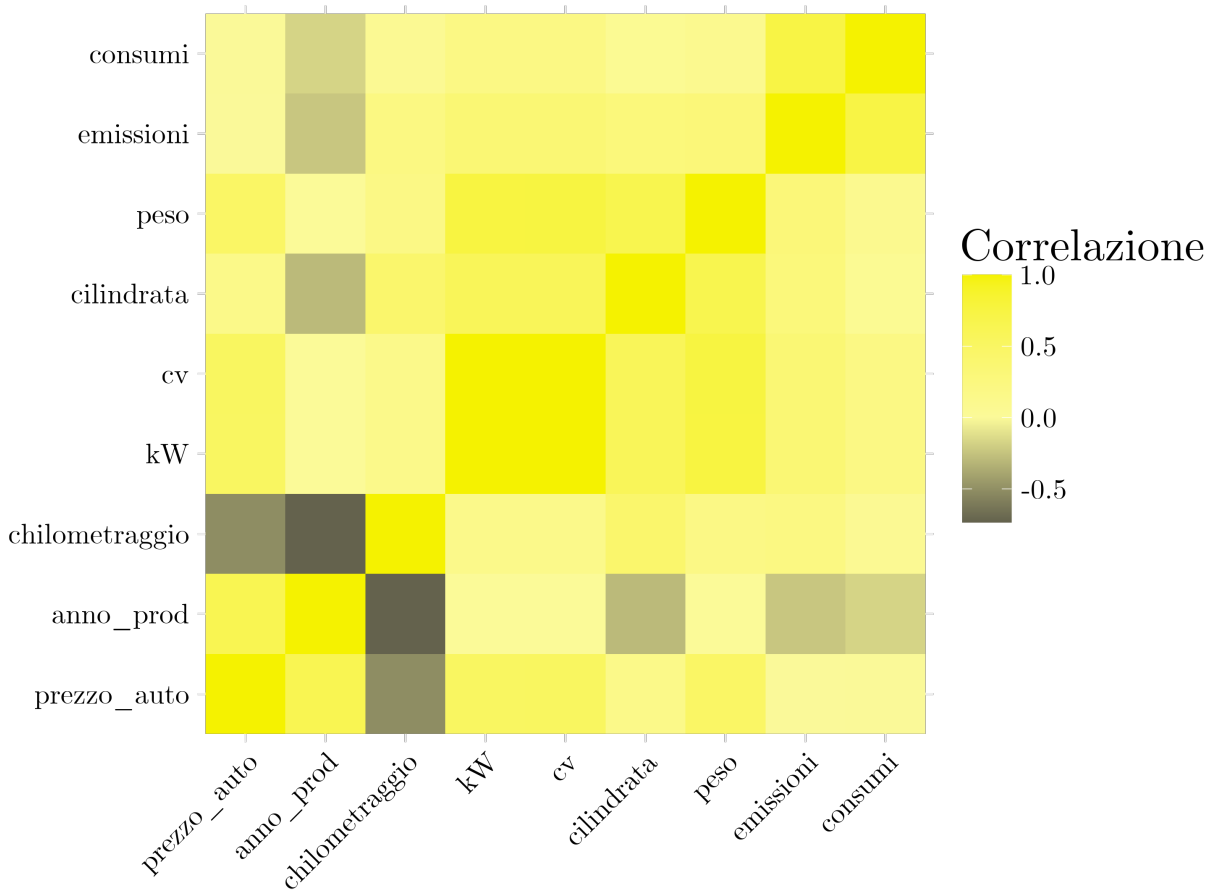
- Relazione considerabile come lineare o quadratica.

Optional più diffusi



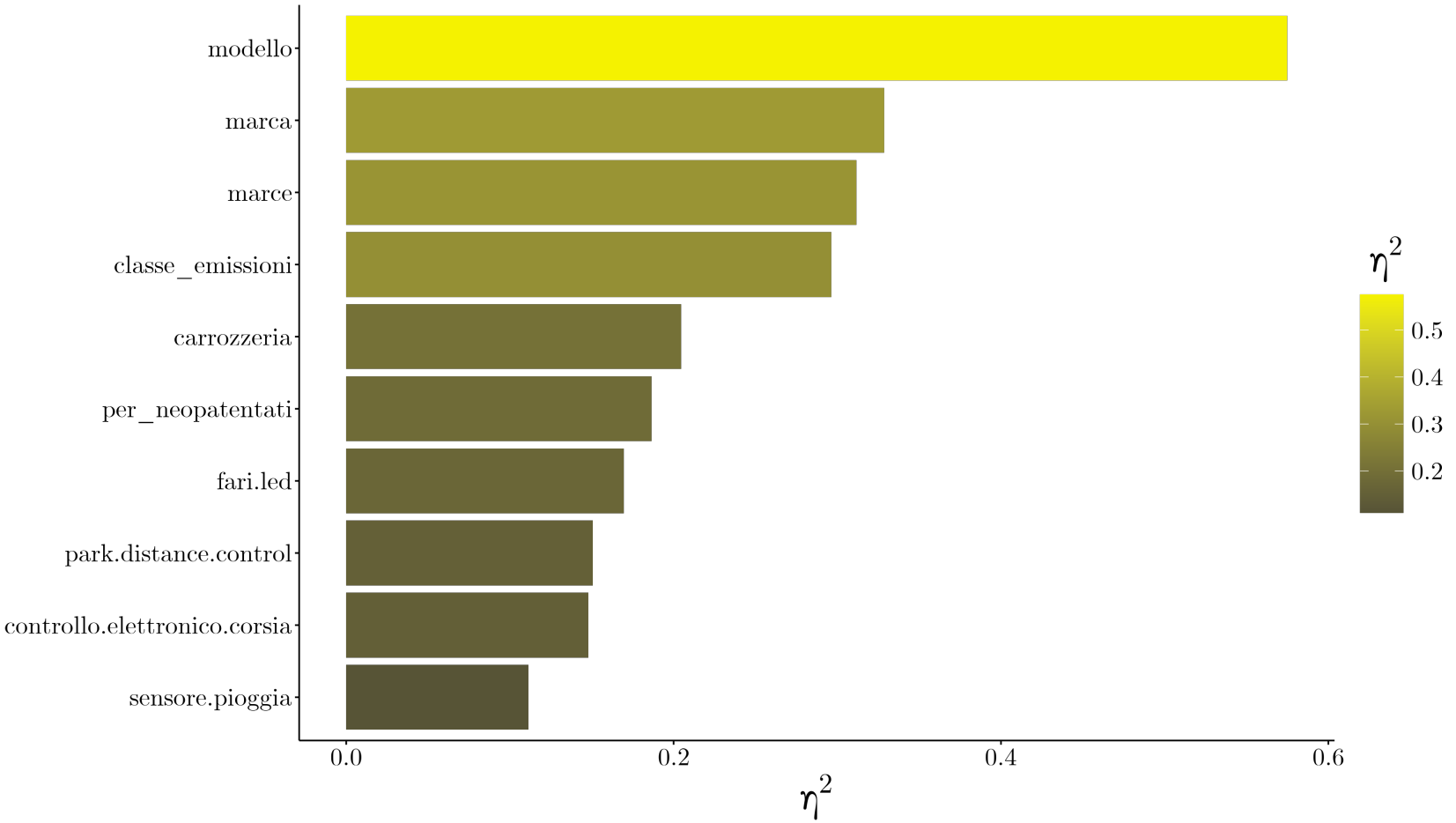
- In molte auto sono presenti il controllo della velocità di crociera e il sistema Isofix per i seggiolini.

Correlazione tra variabili quantitative



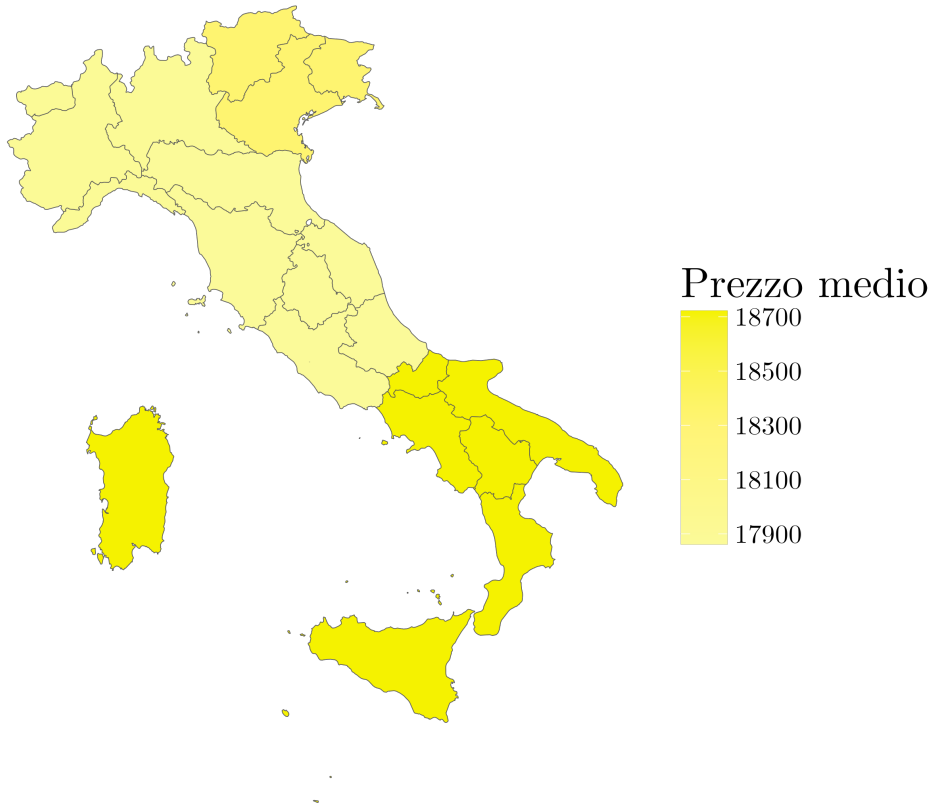
- Le variabili `cv` e `kW` sono una la versione riscalata dell'altra.

Indice η^2: correlazione tra prezzo e variabili qualitative



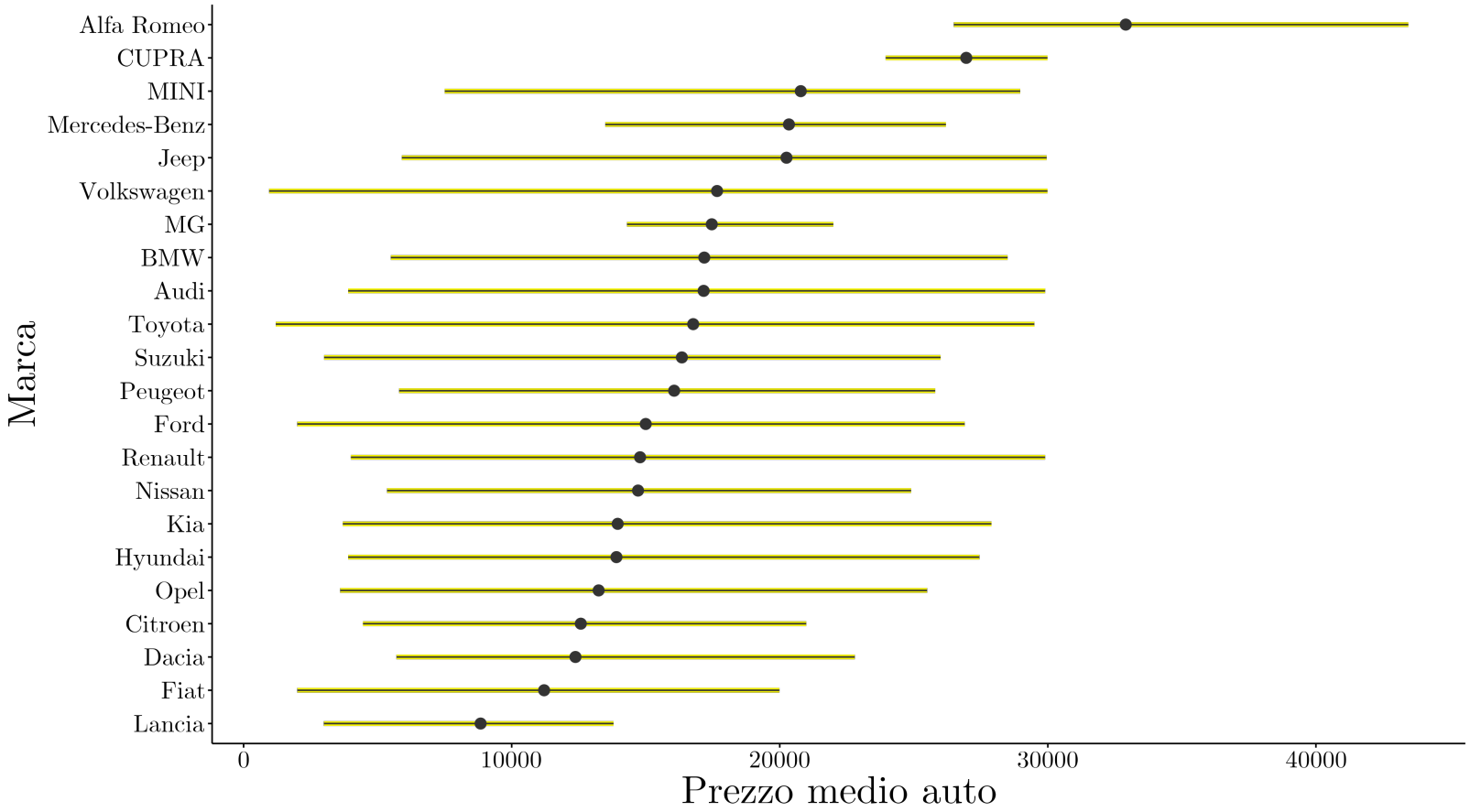
- Naturalmente, modello e marca sono le più correlate con il prezzo.

Prezzo medio per zona geografica



- La differenza di prezzo tra zone geografiche appare minima.

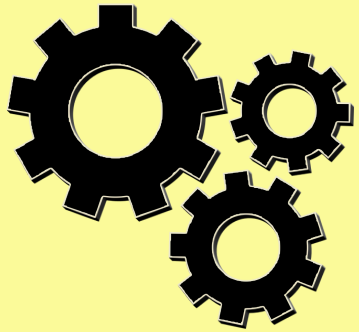
Prezzo medio per marca



- Media calcolata raggruppando i modelli di ogni marca.

Analisi statistiche

modelli, regole associative, clustering



Roadmap

1) Trasformazione delle variabili:
   - rimozione di `kW` (problemi di collinearità con `CV`);
   - `prezzo_auto` $\rightarrow log(\texttt{prezzo\_auto})$
     `chilometraggio` $\rightarrow log(\texttt{chilometraggio}+1)$
     `cilindrata` $\rightarrow log(\texttt{cilindrata})$
     `peso` $\rightarrow log(\texttt{peso})$
2) Regole associative e LMM per vedere se ha senso considerare optional e indicazioni sulla zona geografica.
3) Modelli per la variabile prezzo.
4) Clustering per raggruppare auto simili (variabili nella scala originaria).

Regole associative: perché?

- Le 45 variabili indicatrici inerenti agli optional possono facilmente essere utilizzate per costruire una matrice di “transazioni”.
- Transazione $\equiv$ auto.
- Opportunità...
  - per AutoScout24:
    - verificare la coerenza negli optional elencati dall’inserzionista;
  - per i potenziali acquirenti:
    - avere un miglior quadro della ricchezza della dotazione di un’auto;
  - per le concessionarie:
    - proporre optional o pacchetti di optional mirati su versioni base dei modelli.

Metodo e soglie

Regole associative

- Algoritmo Apriori per trovare le regole associative.
- Si ricorda che, detti A e C antecedente e conseguente della regola,

$$\text{supporto} = \mathbb{P}(C \cap A), \quad \text{fiducia} = \mathbb{P}(C|A) \quad \text{e} \quad \text{lift} = \frac{\mathbb{P}(C|A)}{\mathbb{P}(C)}.$$

- Soglie su fiducia e supporto: 0.9 e 0.5.
- Si selezionano le regole con lift > 1.

Risultati

Regole associative

- Si scoprono diverse regole associative con fiducia pari a 1, tra cui:
  - `{luci diurne}` $\Longrightarrow$ `{luci diurne led}`
  - `{sensori parcheggio}` $\Longrightarrow$ `{park distance control}`
- Le due regole meno banali sono:
  - `{volante multifunzione}` $\Longrightarrow$ `{autoradio}`
  - `{no controllo elettronico corsia}` $\Longrightarrow$ `{no riconoscimento segnali}`

  con fiducia rispettivamente pari a 0.93 e 0.92.
- Le altre regole coinvolgono principalmente gli optional `Android Auto` e `Apple CarPlay` che sono quasi sempre assenti o presenti insieme.

Risultati (2)

Regole associative

- Abbassando la soglia sulla fiducia da 0.9 a 0.7 si trovano risultati più interessanti (regole ordinate per fiducia decrescente):
  - $\{\texttt{park distance control, cruise control}\} \Longrightarrow \{\texttt{cerchi in lega}\}$
  - $\{\texttt{no touch screen}\} \Longrightarrow \{\texttt{no luci diurne led}\}$
  - $\{\texttt{autoradio, cruise control}\} \Longrightarrow \{\texttt{bluetooth}\}$
  - $\{\texttt{volante in pelle}\} \Longrightarrow \{\texttt{isofix}\}$

Effetto della zona geografica sul prezzo

- Il legame tra prezzo e zona geografica dell'auto è statisticamente significativo?
- LMM (stima REML) con intercetta variabile per...
  - regione:

$$ICC = \frac{\sigma_a^2}{\sigma_a^2 + \sigma_\epsilon^2} = 0.04$$

  - zona geografica:

$$ICC = \frac{\sigma_a^2}{\sigma_a^2 + \sigma_\epsilon^2} < 0.01$$

- Le variabili `regione` e `zona_geografica` vengono perciò scartate.

Gradient boosting

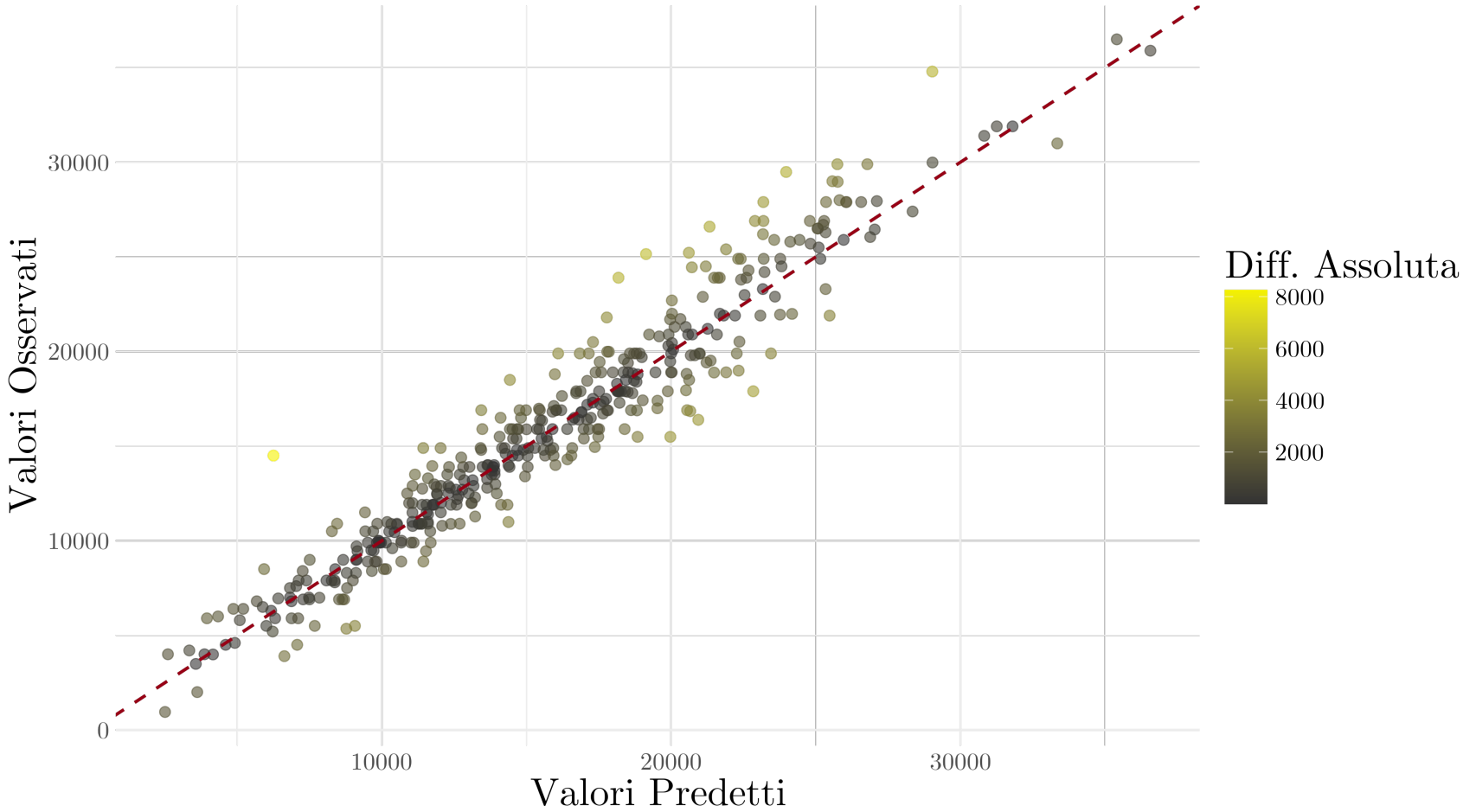
- Divisione semi-casuale del dataset in insieme di stima e verifica.
- Nuova variabile `optional`:
  $\texttt{optional}_{ji} = \sum_i I(\text{optional}_{ji} == 1), \quad j = 1, \ldots, 1591, i = 1, \ldots, 45.$
- Creazione di una griglia equispaziata per i parametri di regolazione:
  - numero di alberi: da 100 a 1000;
  - shrinkage: da 0.02 a 0.1;
  - profondità: da 1 a 5.
- Scelta della combinazione migliore dei parametri in base al MSE.

Numero di alberi	Shrinkage	Profondità	MSE
500	0.06	5	2865538
800	0.04	4	2908419
800	0.06	4	2811461
1000	0.04	4	2882468
1000	0.04	5	2863496

Performance del Gradient boosting

Marca	Modello	Prezzo (€)	Prezzo previsto (€)	Giudizio AS24	Giudizio previsto
Renault	Captur	19.900	18.253	Buono	Pessimo
Toyota	Yaris Cross	11.990	12.911	Buono	Ottimo
Audi	A3	25.900	27.302	Ottimo	Ottimo
BMW	X1	14.900	10.849	Ottimo	Pessimo
Citroen	C3	13.300	11.562	Ottimo	Pessimo
CUPRA	Formentor	27.950	27.680	Ottimo	Buono
Fiat	500	7.900	8.216	Ottimo	Buono
Hyundai	i10	8.400	7.336	Ottimo	Pessimo
Kia	Sportage	10.450	10.481	Ottimo	Buono
Lancia	Ypsilon	4.500	4.792	Ottimo	Buono
Mercedes-Benz	GLA	24.900	22.626	Ottimo	Pessimo
MG	ZS	14.900	15.948	Ottimo	Ottimo
MINI	Countryman	11.400	11.513	Ottimo	Buono
Nissan	Qashqai	24.900	22.331	Ottimo	Pessimo
Opel	Corsa	10.900	11.542	Ottimo	Ottimo
Suzuki	Ignis	15.900	15.104	Ottimo	Pessimo
Dacia	Sandero	10.490	10.335	Super	Buono
Fiat	Panda	10.850	10.654	Super	Buono
Ford	Puma	18.900	17.296	Super	Pessimo
Jeep	Renegade	16.900	16.106	Super	Pessimo
Peugeot	208	9.900	9.962	Super	Buono
Volkswagen	T-Roc	21.990	23.778	Super	Ottimo

Performance del Gradient boosting



Confronto tra modelli

	RMSE	R^2
Gradient boosting (GB)	0.14	0.93
Multivariate Adaptive Regression Spline (MARS)	0.14	0.90
Random Forest (RF)	0.15	0.89
Modello lineare con selezione stepwise	0.16	0.88
Least Angle Regression (LAR)	0.16	0.88
Ridge	0.16	0.88
Projection Pursuit Regression (PPR)	0.16	0.88
Relevance Vector Machines con kernel lineare	0.35	0.88
Rete neurale	0.20	0.79

Analisi dei cluster: metodi

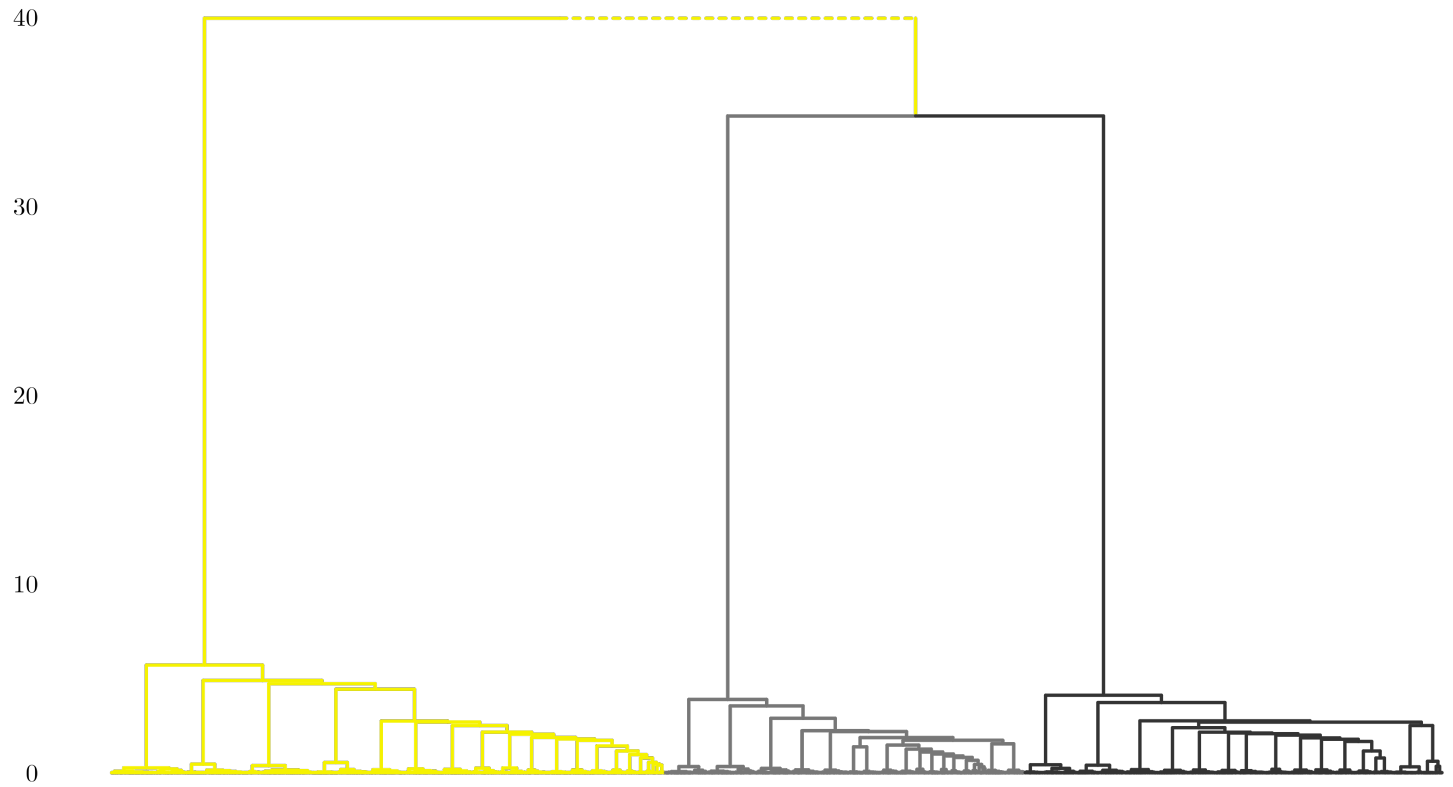
1) Variabili quantitative e qualitative insieme.
2) Solo variabili quantitative:
   a) metodo gerarchico;
   b) metodo non gerarchico (*k*-means);
   c) model-based.

Clustering con variabili miste

- Rimozione di `kW` e `modello`.
- Variabili con $\rho > 0.3$: `anno_produzione`, `chilometraggio`, `cv`, `peso`.
- Variabili con $\eta^2 > 0.3$: `marca`, `marce`.
- Standardizzazione:
  - qualitative trasformate in indicatrici;
  - quantitative riscalate.
- Misura di dissimilarità: indice di Gower.
- Metodo di agglomerazione: legame di Ward.

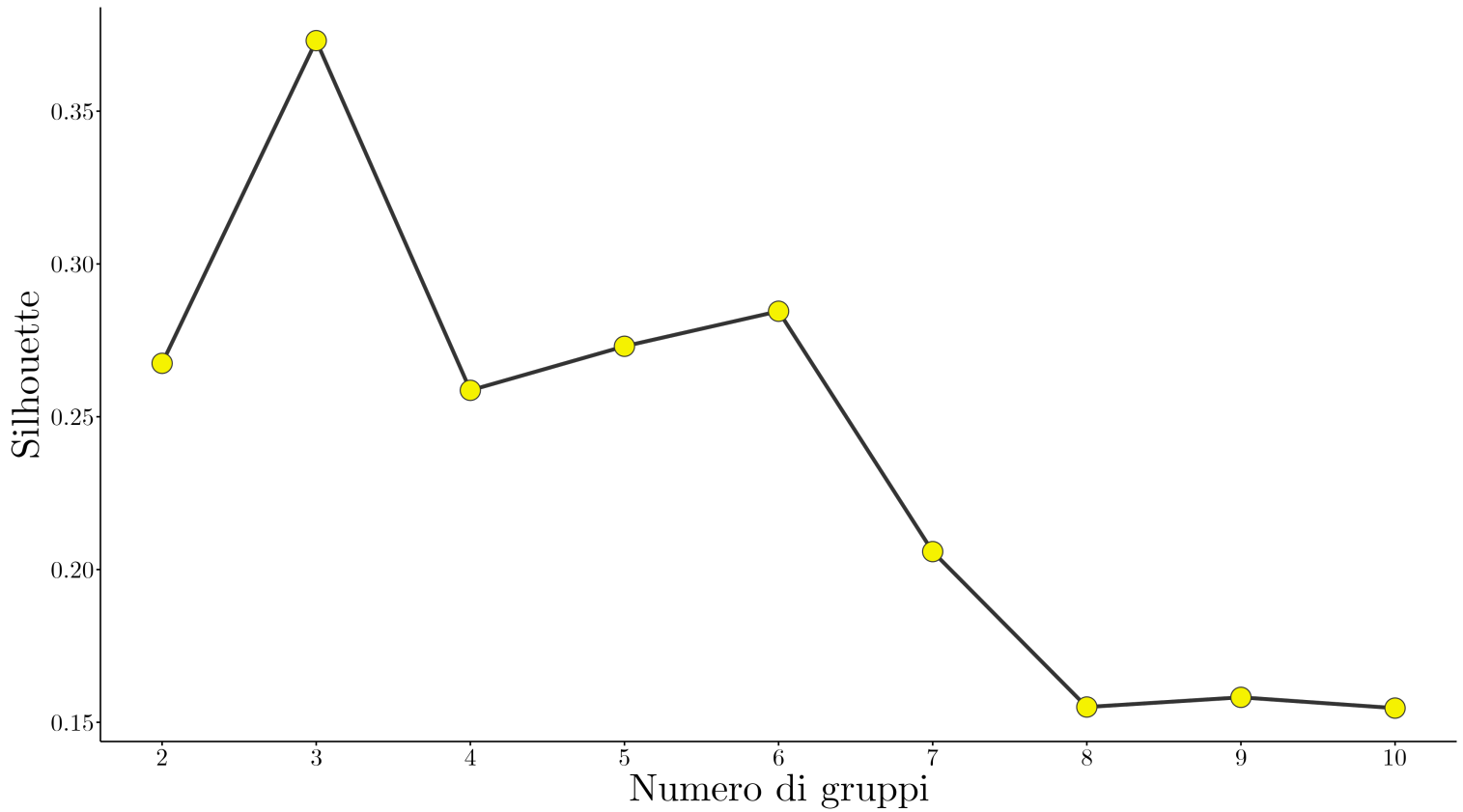
Dendrogramma

- Il dendrogramma evidenzia la presenza di tre gruppi.



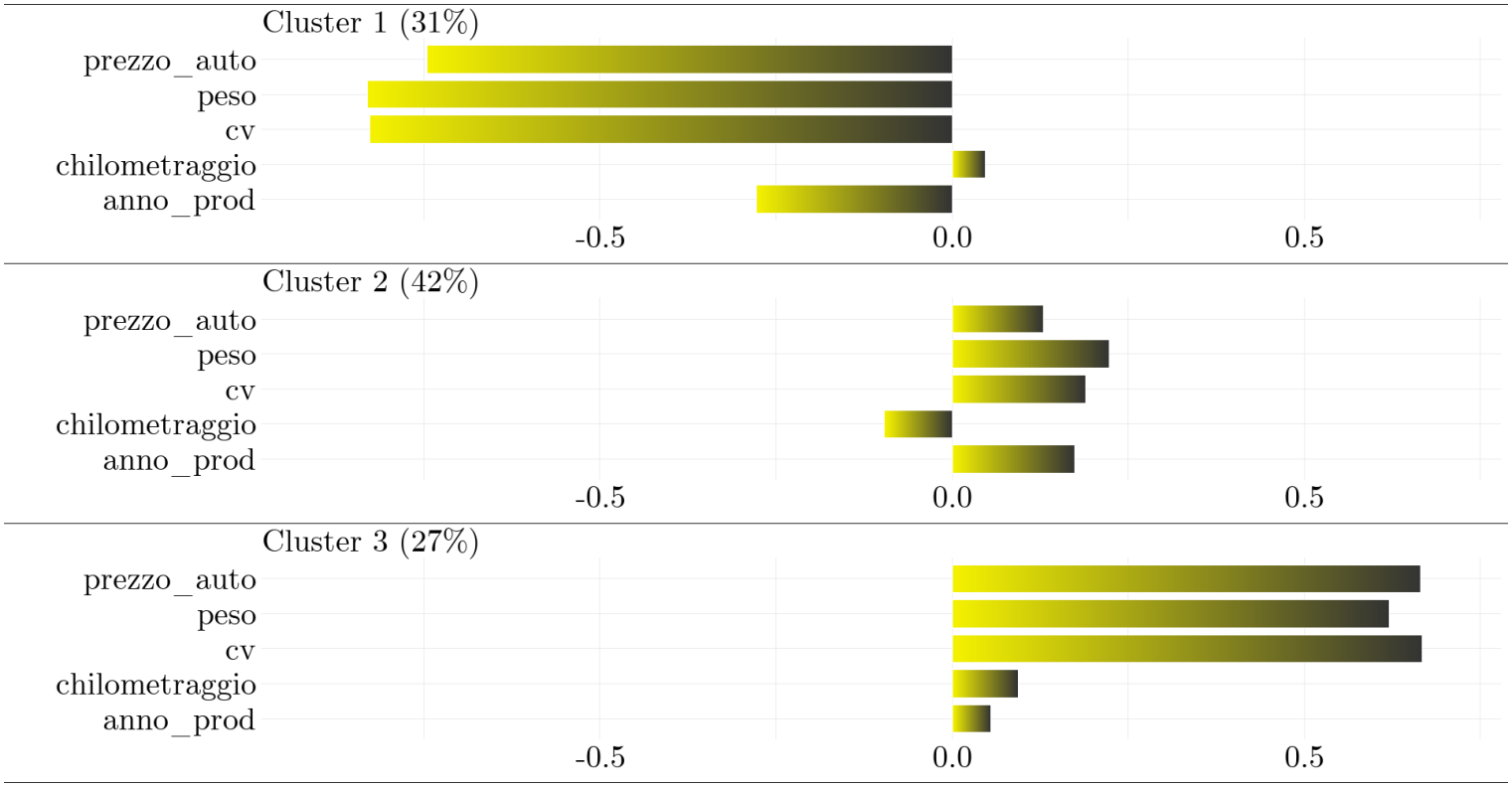
Silhouette

- Silhouette al variare del numero di gruppi.



Caratteristiche dei gruppi

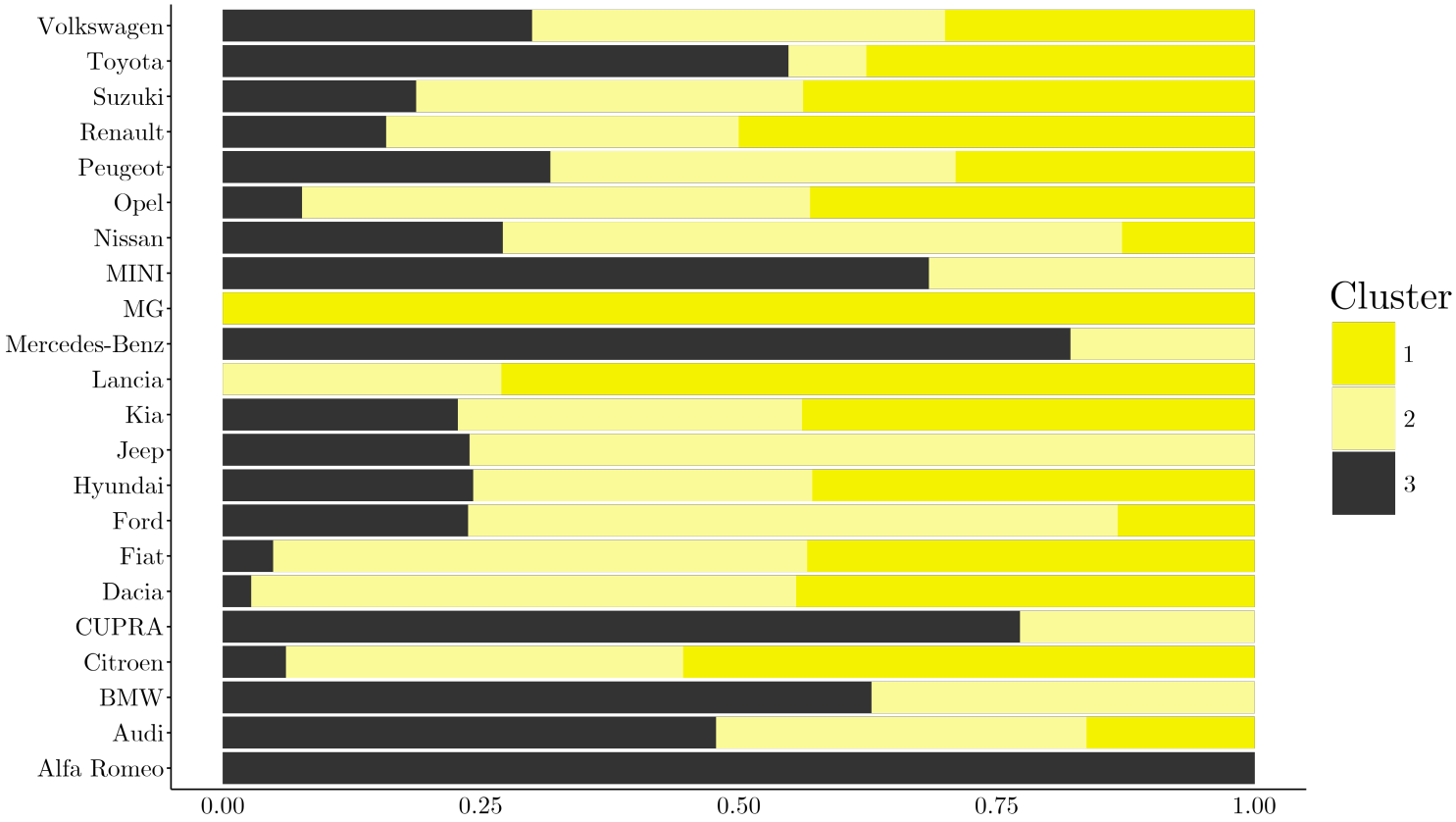
Clustering non gerarchico, variabili di natura mista



- Prevalenza di auto a 5 marce nel cluster 1, di auto con 6 o più marce nel cluster 2 e di auto con il cambio automatico nel cluster 3.

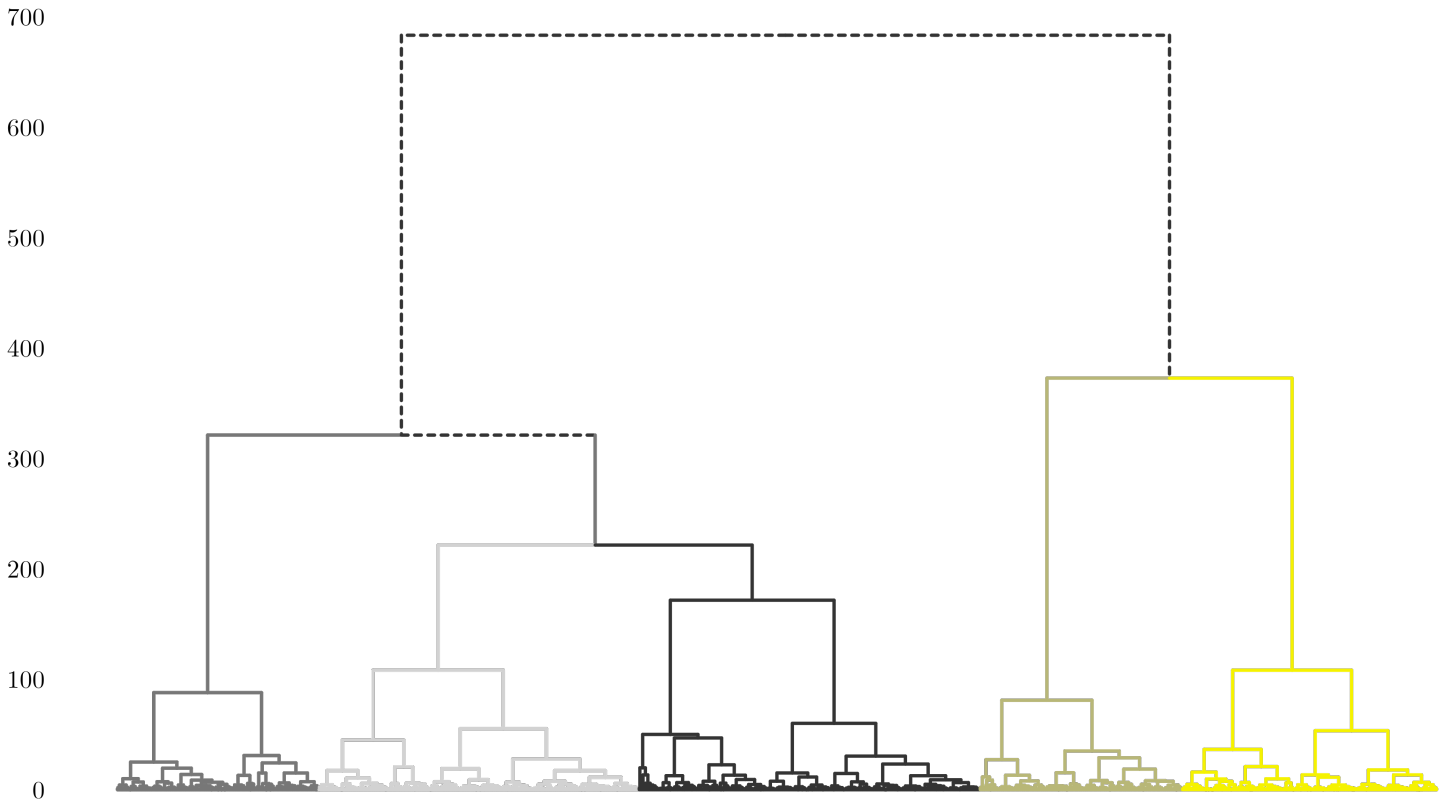
Suddivisione delle marche entro i cluster

- I modelli delle marche MG e Alfa Romeo appartengono tutti ad un unico cluster.



Clustering gerarchico con variabili quantitative

- Si considera ancora la variabile `optional` (totale accessori di un'auto).
- Distanza euclidea, legame di Ward.
- Il dendrogramma suggerisce 5 gruppi.



Caratteristiche dei gruppi

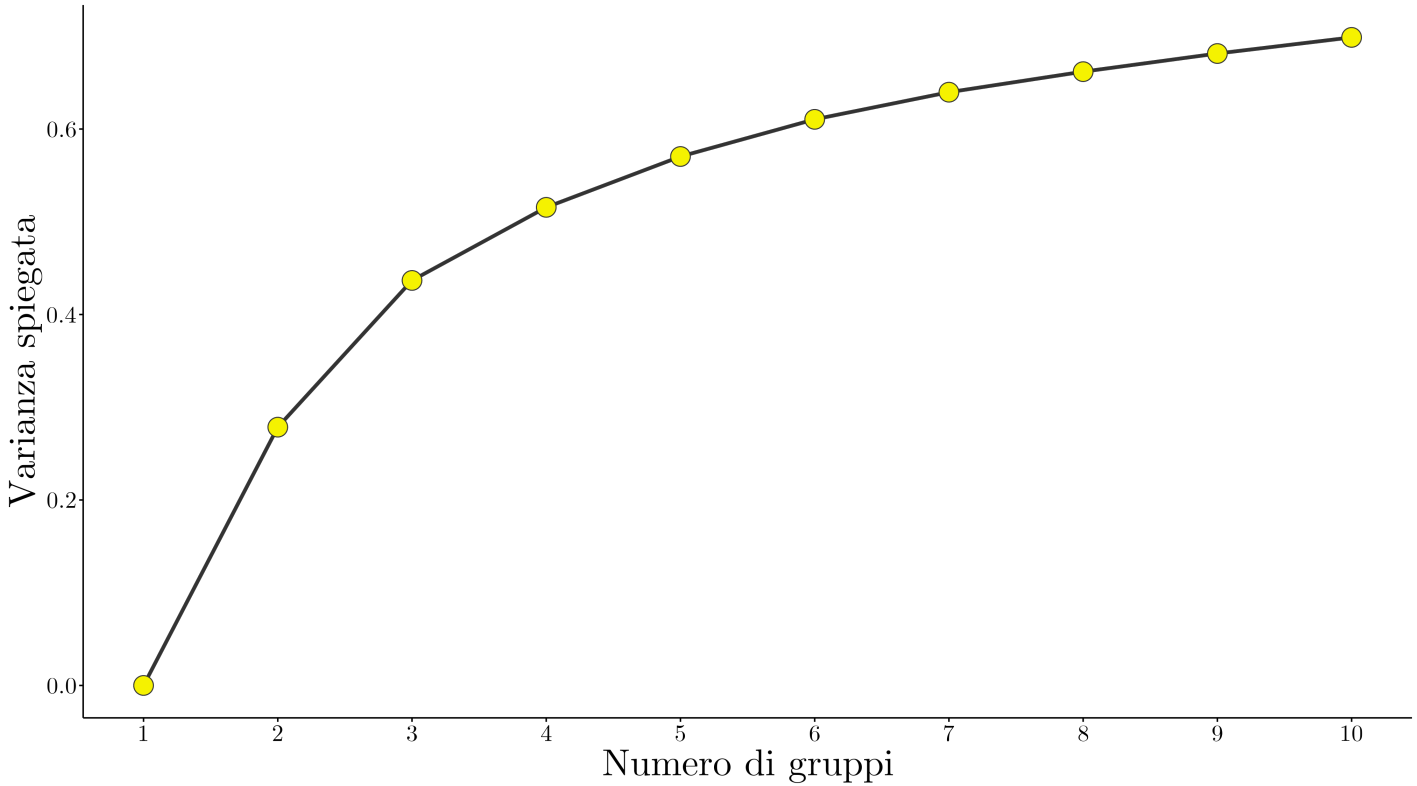
Clustering gerarchico, variabili quantitative

- Cluster 1: utilitarie di recente produzione, non molto potenti e con dotazioni essenziali.
- Cluster 2: auto usate, con alto chilomeraggio, e pochi optional.
- Cluster 3: auto di fascia media relativamente nuove e un prezzo più elevato rispetto alla media.
- Cluster 4: auto recenti, di fascia alta o sportive.
- Cluster 5: auto pesanti, potenti e relativamente economiche, usate ma con dotazione adeguata.

Cluster	prezzo_auto	anno_prod	chilometraggio	cv	peso	optional
1	↓	↑	↓	↓	↓	–
2	↓	↓	↑	↓	↓	↓
3	↑	↑	↓	–	–	↓
4	↑	↑	↓	↑	↑	↑
5	↓	↓	↑	↑	↑	↑

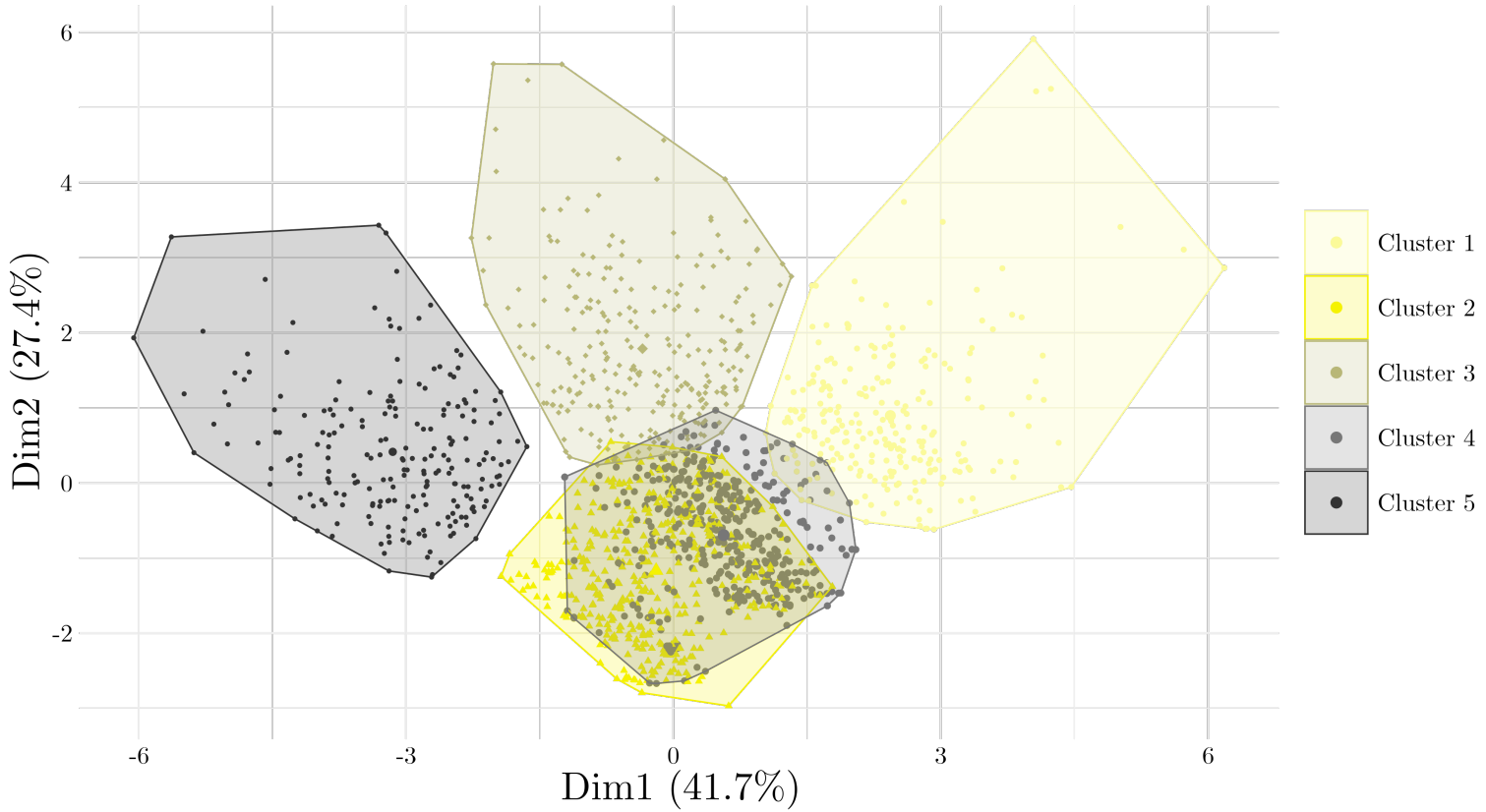
Algoritmo *k*-means e varianza spiegata

- Scree plot varianza spiegata per 1, . . . , 10 cluster.



- Grafico poco informativo. Si considerano 5 cluster sfruttando il risultato ottenuto con il clustering gerarchico.

Rappresentazione grafica dei cluster

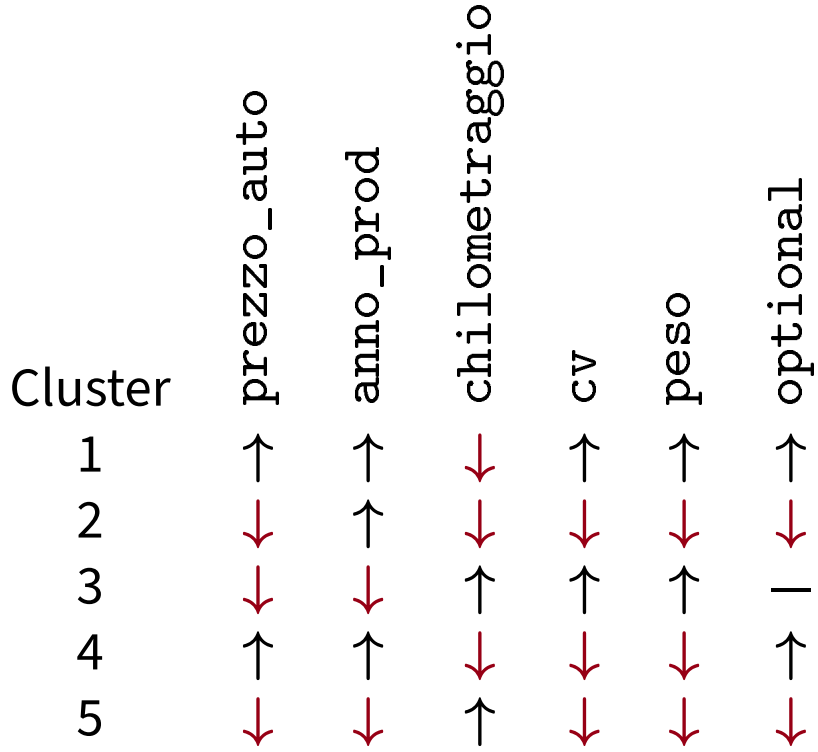


- Grande sovrapposizione tra i cluster 2 e 4.

Caratteristiche dei gruppi

Clustering non gerarchico

- Cluster 1: auto di alta gamma: nuovi SUV o auto sportive.
- Cluster 2: auto economiche, non molto potenti, con dotazione basica ma di recente produzione.
- Cluster 3: SUV o auto di fascia media usate.
- Cluster 4: auto berline con un focus maggiore sui comfort piuttosto che sulle prestazioni.
- Cluster 5: auto usate non potenti e con dotazioni base.

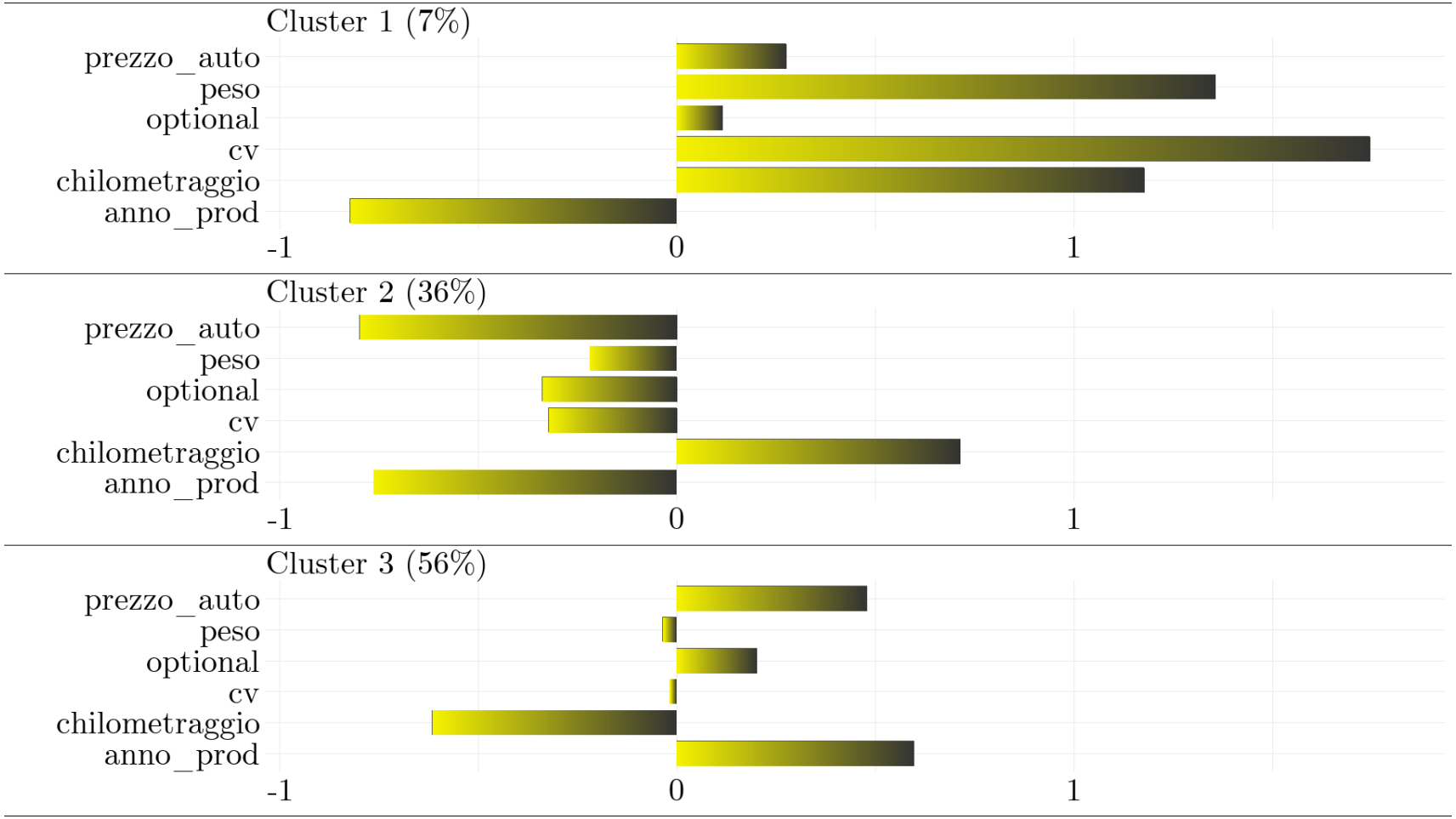


Clustering model-based

```
----------------------------------------------------
Gaussian finite mixture model fitted by EM algorithm
----------------------------------------------------

Mclust VVV (ellipsoidal, varying volume, shape, and orientation) model with 5 components:

log-likelihood    n  df       BIC      ICL
     -8184.647 1591 139 -17394.02 -17681.1

Clustering table:
  1   2   3   4   5
408 139 757 115 172


----------------------------------------------------
Gaussian finite mixture model fitted by EM algorithm
----------------------------------------------------

Mclust VEV (ellipsoidal, equal shape) model with 3 components:

 log-likelihood    n df       BIC       ICL
      -9723.349 1591 73 -19984.86 -20356.76

Clustering table:
  1   2   3
116 579 896
```

Caratteristiche dei gruppi

Clustering model-based



Conclusioni

e risposta alla domanda di business

Conclusioni e suggerimenti per l'azienda

Optional

- Gli inserzionisti tendono:
  - ad inserire elenchi incompleti di optional;
  - a ripeterne alcuni con nomi diversi;
  - a non specificare tutti gli optional per le macchine costose.
- 💡 Gli optional potrebbero essere utili a migliorare i modelli, ma è meglio ottenere il loro elenco da siti specializzati piuttosto che fidarsi di ciò che inseriscono gli utenti.

Conclusioni e suggerimenti per l'azienda

Caratteristiche delle auto

- Anche le informazioni sulle caratteristiche dell'auto possono essere:
  - poco congruenti (es. auto con 0 cilindri);
  - mancanti.
- Ancora, il consiglio è far affidamento ai dati di siti specializzati per associare le informazioni corrette ad ogni vettura nel database.
- In definitiva, si suggerisce la costruzione manuale di un dataset con le caratteristiche ufficiali di quanti più modelli possibile.
- Questo eviterebbe anche di ricorrere a tecniche di imputazione.

Conclusioni e suggerimenti per l'azienda

Metodologia

- La zona geografica di provenienza dell'auto non ha una grande influenza sul prezzo.
- Tra i modelli di previsione del prezzo, il gradient boosting è quello che dà i risultati migliori.
- Le procedure di clustering possono essere utili per:
  - implementare un sistema di raccomandazione delle auto in base alla loro appartenenza nei cluster;
  - creare una sezione di auto suddivise per cluster, cioè per caratteristiche simili.

Quanto vale la mia auto?

- È stata presa come riferimento una Skoda Kamiq Scoutline (la mia!) con le caratteristiche indicate nella tabella sottostante, e ne è stato stimato il prezzo attraverso il gradient boosting.

#	Variabile	Valore
1	`marca`	Skoda
2	`modello`	Kamiq
3	`regione`	Toscana
4	`zona_geog.`	Centro
6	`anno_prod`	2019
7	`carrozzeria`	SUV
9	`trazione`	Anteriore
10	`posti`	5
11	`porte`	5
13	`chilometraggio`	88256
15	`cv`	90
16	`marce`	Automatico
20	`carburante`	Benzina
21	`classe_emissioni`	Euro 6 Plus
28	`finiture`	Grigie
31	`consumi`	5.6 l/100Km

La previsione del prezzo di questa auto è **13613.59€**.

Riferimenti e link al progetto GitHub

- Wickham, H., & Wickham, M. H. (2016). Package 'rvest'. URL: https://cran.r-project.org/web/packages/rvest/rvest. PDF, p156.
- Bassi, Francesca. "Statistica per Analisi Di Mercato. Metodi e Strumenti." Statistica per Analisi Di Mercato: Metodi e Strumenti, Pearson, 2022.
- Hastie, Trevor J., Tibshirani, Robert & Friedman, Jerome, "The elements of statistical learning: data mining, inference, and prediction." New York: Springer, 2009.

- Link a codice R e dataset:

Grazie per l'attenzione!