# Caso di Studio: Gestione Intelligente di Sistemi di Vendita di Automobili

Corso di Ingegneria della Conoscenza
Università degli Studi di Bari

Gianmarco Rutigliano 747002
g.rutigliano30@studenti.uniba.it
https://github.com/Gyanma-rev/Car_Sales

November 5, 2024

# Contents

# 1 Introduzione

Il presente caso di studio ha come sfera di interesse il mercato di compravendita di automobili usate. Il dataset utilizzato come fonte primaria di conoscenza è il dataset "Used Car Auction Prices", pubblicato su Kaggle [1]. Il dataset raccoglie 558,837 istanze di vendite di automobili dal 2013 al 2015 negli Stati Uniti. Per ognuna di queste raccoglie vari dati inerenti l'automobile, quali modello, anno di produzione o valutazione sulle condizioni, e la transazione, quali il prezzo di vendita, il prezzo di mercato dell'auto o la data della vendita. L'elaborato si compone di molteplici parti.

- La prima parte concerne le operazioni di preprocessing e lavorazione del dataset: il dataset viene esaminato per rimuovere dati incompleti o poco utili e acquisire informazioni iniziali sui dati utili per direzionare le fasi successive e progettare ulteriori adattamenti.
- La seconda parte concerne la costruzione di una Knowledge Base basata sul dataset: questo permette di ottenere informazioni aggiuntive sui dati tramite la definizione di regole.
- La terza parte concerne la definizione di un modello di apprendimento non supervisionato di clustering con K-Means tra i dati delle automobili, con focus particolare sul confronto di possibili modelli e loro caratteristiche.

## 1.1 Elenco argomenti di interesse

- Rappresentazione di conoscenza e ragionamento relazionale: utilizzo di Prolog per il ragionamento su una base di conoscenza.
- Apprendimento non supervisionato: design di un modello K-Means di hard clustering basato sui dati.

# 2 Preprocessing and Data Preparation

## 2.1 Descrizione dei Dati

I dati presentano le seguenti feature:

**Feature Categoriche:**

- vin: *string, len = 17*; Vehicle Identification Number, un identificatore univoco per ogni veicolo;
- make: *string*; indica il nome del produttore del veicolo;
- model: *string*; indica il modello specifico del veicolo;
- trim: *string*; indica una classificazione aggiuntiva del veicolo;
- body: *string*; rappresenta una classificazione specifica del veicolo, che differenzia elementi quali forma o dimensioni (es. SUV, Sedan, Coupe, etc);
- transmission: {*automatic, manual*}; indica la tipologia della trasmissione del veicolo;
- state: *string*; indica lo stato, all'interno di USA o Canada, in cui avviene l'asta di vendita;
- color: *string*; indica il colore della carrozzeria del veicolo;
- interior: *string*; indica il colore degli interni del veicolo;
- seller: *string*; indica il nome dell'entità che ha venduto il veicolo;
- saledate: *string*; indica la data di vendita del veicolo: a causa della formattazione dei dati, la feature è da considerarsi in questa fase Categorica, tuttavia si prevede di effettuare delle elaborazioni in modo da renderla quantitativa;

**Feature Quantitative:**

- year: *integer*; indica l'anno del modello del veicolo;

- condition: *float, compreso tra 1.0 e 5.0*; indica, in base ad una scala, le condizioni di funzionamento del veicolo;
- odometer: *integer*; indica la distanza in miglia percorsa dall'automobile al momento della vendita;
- mmr: *integer*; indica il valore in USD stimato di mercato del veicolo secondo l'indice Manheim Market Report [2];
- sellingprice: *integer*; indica il prezzo in USD a cui il veicolo è stato venduto;

## 2.2 Qualità dei dati

**Problemi di formattazione** Per estrarre i dati dal CSV, è stato necessario sostituire le virgole (,) nei valori. Questo ha interessato in particolare la colonna 'trim', in cui il valore Connectivity, Navitgation" è stato sostituito con "Connectivity and Navigation".

### 2.2.1 Dati Mancanti

Il dataset presenta diverse istanze con dati mancanti.
Di seguito si riportano le feature che presentano dati mancanti e l'occorrenza e la quantità di feature mancanti per istanza.

| Feature | Istanze | Percentuale |
|---|---|---|
| transmission | 65,357 | 11.7% |
| body | 13,195 | 2.3% |
| condition | 11,794 | 2.1% |
| trim | 10,651 | 1,9% |
| model | 10,399 | 1,9% |
| make | 10,301 | 1,8% |
| color | 749 | 0,1% |
| interior | 749 | 0,1% |
| odometer | 94 | 0,01% |

Table 1: Numero di istanze aventi il valore di una specifica feature mancante.

| Numero di Feature Mancanti | Istanze |
|---|---|
| 1 | 72,631 |
| 2 | 3,156 |
| 3 | 313 |
| 4 | 8,527 |
| 5 | 1,820 |
| 6 | 27 |
| 7 | 3 |
| 8 | 2 |
| Totale | 86480 |

Table 2: Numero di istanze aventi il valore di n feature mancante.

Si può notare che il numero di istanze con valori mancanti è significativo: rappresenta infatti il 15.5% del totale delle istanze. Tuttavia, il numero complessivo di istanze, di 558,837 o 472,358 escludendo quelle incomplete, è già molto consistente e permette un'analisi adeguata al task, quindi rimuovere queste istanze non rappresenta necessariamente una perdita di informazione notevolmente peggiorativa. Per questa ragione si è deciso di rimuovere le istanze incomplete.

### 2.2.2 Dati Duplicati

Il dataset non presenta dati duplicati.

### 2.2.3 Adeguamento delle feature ai fini del task

- Unificazione di valori:

  La feature 'Body' presenta coppie di valori che rappresentano lo stesso concetto ma sono espressi con formattazione diversa, ad esempio "Regular Cab" e "regular-cab", oppure "Sedan" e "sedan". Per rimuovere queste ambiguità, e al contempo preparare il dataset alla creazione della base di conoscenza, si sono applicate delle modifiche ai dati: in particolare si è deciso di

  - portare tutti i dati delle variabili categoriche in carattere minuscolo
  - sostituire gli spazi con trattini ( - )

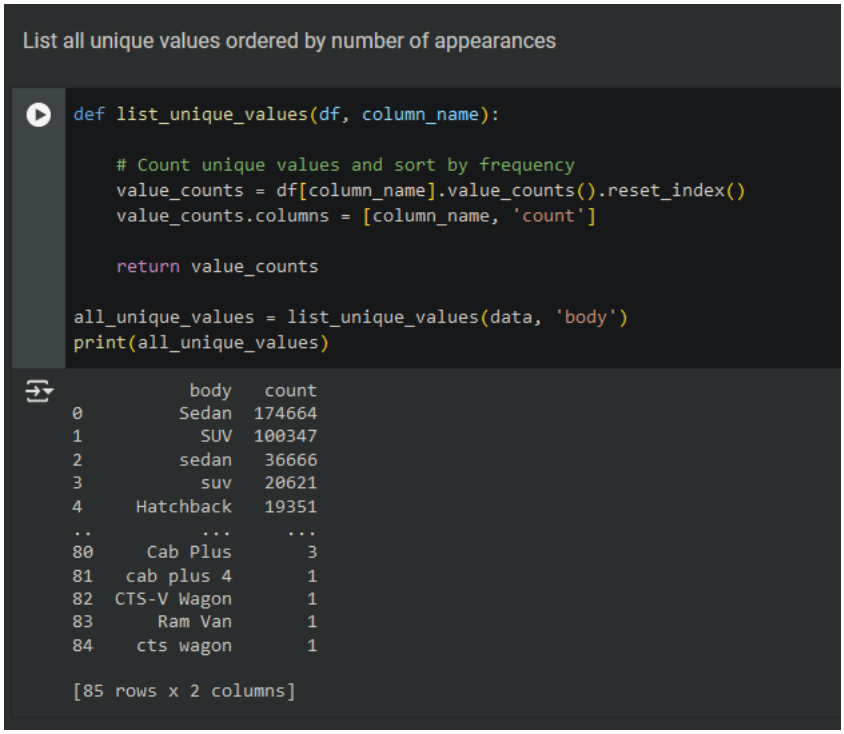


Figure 1: Conteggio valori unici di 'body'

  - rimuovere apici ( ' ) e doppi apici ( " )

- Sostituzione di valori:

  Le feature 'Color' e 'Interior' presentano un valore illeggibile. Si è deciso di sostituire il valore con il valore sconosciuto: "unknown".

## 2.3 Esplorazione dei Dati

Di seguito vengono inseriti dati notevoli riguardanti diverse feature. Per agevolare la lettura, i dati sulla feature 'year' sono stati calcolati prima di effettuare la conversione descritta in precedenza.

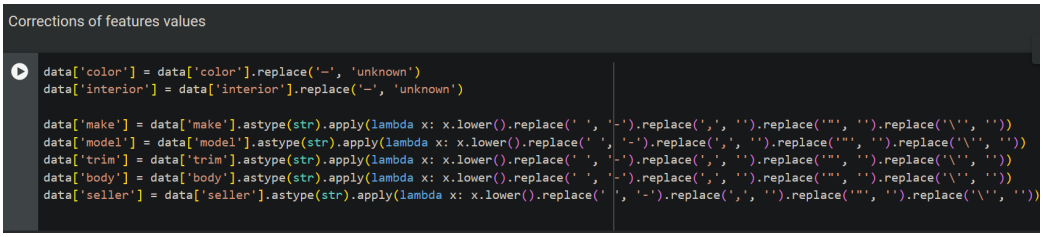


Figure 2: Enter Caption

| Feature | Max | Min | Mode | Median | Mean | StdDev | UniqueVal |
|---|---|---|---|---|---|---|---|
| year | 2015 | 1990 | 2013 | 2012 | 2010.21 | 3.82 | 26 |
| condition | 5 | 1 | 1.9 | 3.6 | 3.42 | 0.94 | 41 |
| odometer | 999,999 | 1 | 1 | 51,081 | 66,698.39 | 51,939.47 | 160,429 |
| mmr | 182,000 | 25 | 11,650 | 11,650 | 1,230 | 9531.91 | 1,099 |
| sellingprice | 230,000 | 1 | 12,000 | 12,000 | 13,690.34 | 9,612.76 | 1,086 |

Table 3: Misure statistiche delle variabili quantitative.

| Feature | make | model | trim | body |
|---|---|---|---|---|
| 1. | Ford | Altima | Base | Sedan |
| 2. | Chevrolet | Fusion | SE | SUV |
| 3. | Nissan | F-150 | LX | Hatchback |
| 4. | Toyota | Camry | Limited | Minivan |
| 5. | Dodge | Escape | LT | Coupe |
| N° valori diversi | 53 | 768 | 1509 | 44 |

Table 4: Valori più comuni delle variabili categoriche.

| Feature | transmission | state | color | interior | seller |
|---|---|---|---|---|---|
| 1. | automatic | fl | black | black | ford motor credit company,llc |
| 2. | manual | ca | white | gray | the hertz corporation |
| 3. | - | tx | silver | beige | nissan-infiniti lt |
| 4. | - | ga | gray | tan | santander consumer |
| 5. | - | pa | blue | unknown | avis corporation |
| N° valori diversi | 2 | 34 | 20 | 17 | 11927 |

Table 5: Valori più comuni delle variabili categoriche.

## 2.4 Feature Aggiuntive

### 2.4.1 Identification

Per identificare correttamente le transazioni, è stata aggiunta la colonna 'Transaction_ID", che funge da chiave primaria del dataset. Infatti l'identificativo 'vin' già presente è associato esclusivamente alle automobili: la combinazione di feature che potrebbe rappresentare una chiave primaria sarebbe la coppia 'vin'-'datetime', ma per comodità di approccio alle operazioni di calcolo, si è preferito aggiungere un identificativo.

# 3 Knowledge Base Construction

**Design** In questa sezione viene descritta la costruzione di una knowledge base in Prolog, realizzata con Python utilizzando la libreria *pyswip*. La base di conoscenza è stata successivamente popolata con fatti a partire dal dataset; le regole sono state progettate in modo da permettere di estrarre conoscenza inerente collegamenti e similitudini tra automobili differenti.

**Preprocessing** Per permettere di effettuare operazioni e confronti con le feature 'Year' e 'Saledate', si è deciso di convertirli secondo una scala unificata. Per ogni valore, si è utilizzata la libreria *datetime* per calcolare lo scarto di giorni tra il valore e una data simbolica (01/01/1970), in modo da ottenere un valore comparabile universalmente. La perdita di informazione dovuta alla maggior specificità di 'Saledate', la quale comprende anche l'orario, si ritiene trascurabile.

Feature Aggiuntiva: Viene aggiunta la variabile 'Recent' . Questa rappresenta un valore binario che assume valore *true* se il dato descrive l'asta più recente cronologicamente che ha coinvolto l'automobile in vendita, e *false* altrimenti. Questa ulteriore variabile permette di estrarre il dato più recente sull'automobile e sarà utilizzato nella composizione di alcune regole.

**Individui** Il design della Knowledge Base ruota attorno a due concetti: automobili e aste, rappresentate dai termini Vin e Auction. Le feature del dataset possono essere raccolte in 3 categorie:

- dati *inerenti specificamente l'asta*,
- dati inerenti l'automobile *circoscritti al momento dell'asta*, che possono quindi variare nel tempo
- dati inerenti l'automobile *immutabili nel tempo*

I fatti raccolti nella Knowledge Base sono dunque stati composti assegnando al predicato 'car' tutti i dati dell'automobile che non dipendono dal momento in cui sono raccolti e al predicato 'auction' i dati inerenti l'asta e i dati relativi all'automobile come registrati al momento dell'asta. È stato infine inserito il predicato 'car_sold', che permettere di connettere i dati degli altri predicati. In particolare i fatti sono così composti:

- car(Vin, Make, Model, Year, Body, Transmission)
- auction(Auction, Seller, State, Price, Date, Condition, Odometer, Color, Interior, MMR, Recent)
- car_sold(Auction, Vin)

Le variabili riflettono le feature presenti nel dataset.

**Regole** Sono state composte le seguenti regole per la Knowledge Base:

- was_sold_repeatedly(Car): vera se esistono almeno due aste di vendita dell'auto.
- was_sold_within_one_year(Car): vera se esiste un'asta di vendita dell'auto con data di vendita maggiore dell'anno di produzione per non più di un anno.

- does_color_match_interior(Car): vera se il colore dell'auto è lo stesso degli interni dell'auto.
- was_sold_above_mmr(Car): vera se l'auto è stata venduta ad un prezzo superiore al prezzo di mercato.
- was_sold_below_mmr(Car): vera se l'auto è stata venduta ad un prezzo inferiore al prezzo di mercato.
- was_sold_at_mmr(Car): vera se l'auto è stata venduta ad un prezzo pari al prezzo di mercato.
- high_volume_seller(Seller): vera se esistono almeno 200 aste aventi Seller come venditore
- highly_traded_make(Make): vera se esistono almeno 1000 aste di auto aventi Make come produttore.

- most_recent_sale(Car, Auction): restituisce le aste sull'auto più recenti.
- current_car_color(Car, Color): restituisce il colore dell'auto relativo all'asta più recente.
- current_car_interior(Car, Interior): restituisce il colore degli interni dell'auto relativo all'asta più recente.
- current_car_mmr(Car, MMR): restituisce il MMR dell'auto relativo all'asta più recente.
- current_car_selling_price(Car, Price): restituisce il prezzo dell'auto relativo all'asta più recente.
- current_car_condition(Car, Condition): restituisce la condizione dell'auto relativa all'asta più recente.
- current_car_odometer(Car, Odometer): restituisce il contachilometri dell'auto relativo all'asta più recente.
- last_seller(Car, Seller): restituisce il venditore dell'auto relativo all'asta più recente.

- is_same_color(Car1, Car2): vera se le auto hanno lo stesso colore.
- is_same_interior(Car1, Car2): vera se le auto hanno lo stesso colore di interni.
- is_same_market_price(Car1, Car2): vera se le auto hanno lo stesso prezzo di mercato.
- is_same_selling_price(Car1, Car2): vera se le auto hanno lo stesso prezzo di vendita.
- is_same_condition(Car1, Car2): vera se le auto sono nella stessa condizione.
- is_same_odometer(Car1, Car2): vera se le auto hanno lo stesso contachilometri.
- is_same_make(Car1, Car2): vera se le auto sono dello stesso produttore.
- is_same_production_year(Car1, Car2): vera se le auto sono dello stesso anno di produzione.
- is_same_body(Car1, Car2): vera se le auto sono dello stesso *body*.
- is_same_transmission(Car1, Car2): vera se le auto hanno la stessa trasmissione.

# 4 Clustering

Questa sezione descrive il processo di costruzione di modelli di clustering *k-means*, una tecnica di apprendimento non supervisionato che raggruppa i dati in cluster distinti sulla base della loro vicinanza all'interno dello spazio delle caratteristiche, assegnando ogni elemento al cluster con il centroide più vicino. Senza la necessità di etichette predefinite, questo approccio è particolarmente utile per esplorare e scoprire strutture nascoste nei dati.

**Preprocessing** Scelta di feature: Per il clustering, si è scelto di escludere le feature che non forniscono direttamente dati sull'automobile: in particolare, le feature escluse sono 'vin', 'state', 'saledate', 'transaction_ID', 'most_recent'. Inoltre, per mantenere solo i dati più aggiornati, il dataset è stato filtrato sui dati più recenti sulle automobili

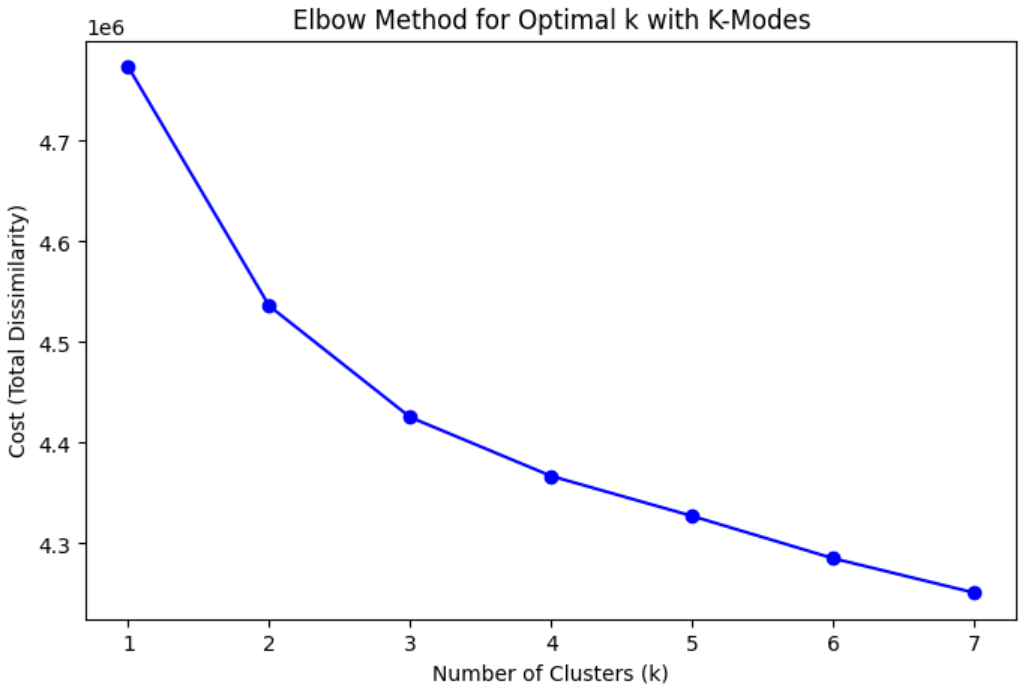


Figure 3: Elbow Method

## 4.1 Elbow Method

Per stimare il numero di cluster più appropriato da individuare, è stato implementato il cosiddetto elbow method. Esso consiste nel calcolare i costi di diversi modelli con un numero crescente di cluster e trovare sulla funzione dei costi il numero più adeguato. Come si può notare, la riduzione dei costi più significativa è individuabile con k = 2, tuttavia si è deciso di effettuare un confronto più completo considerando anche i modelli che costruiscono 3 e 4 cluster.

## 4.2 Analisi dei modelli

L'analisi dei modelli procede esaminando 3 configurazioni di modelli di clustering basati sull'algoritmo K-Means, variando il numero di cluster. Questa analisi permette di osservare come diverse impostazioni influenzino i risultati del clustering e aiutino a rilevare eventuali differenze tra i gruppi formati. I modelli sono stati costruiti con lo strumento Weka, utile perché perme-

```
kMeans
======

Number of iterations: 8
Within cluster sum of squared errors: 1584390.038480827

Initial starting points (random):

Cluster 0: chevrolet,tahoe,2000,suv,automatic,onemain-rem/ramos-autos-&-trucks-inc,ga,2700,1.9,173327,brown,brown,2175
Cluster 1: nissan,pathfinder,2007,suv,automatic,santander-consumer,mo,5000,3.3,153081,blue,gray,5100

Missing values globally replaced with mean/mode

Final cluster centroids:
                                                                  Cluster#
Attribute                                              Full Data         0                  1
                                                      (307428.0)  (130824.0)         (176604.0)
=============================================================================================================
make                                                        ford   chevrolet               ford
model                                                     altima      impala             altima
year                                                  2010.2173  2008.7681          2011.2908
body                                                      sedan       sedan              sedan
transmission                                          automatic   automatic          automatic
seller                                  nissan-infiniti-lt the-hertz-corporation  nissan-infiniti-lt
state                                                       fl          ca                 fl
sellingprice                                         13698.6613  10588.2853         16002.7532
condition                                               3.4293       3.071             3.6946
odometer                                             66650.3988  85930.5799          52368.104
color                                                      black       white              black
interior                                                   black        gray              black
mmr                                                  13840.5897  10802.0461         16091.4699
```

Figure 4: Two Clusters Model

tte di definire modelli di K-Means anche su dataset che contengono feature categoriche.

Esaminando i modelli si possono riconoscere diverse tendenze. Rispetto alle variabili quantitative, i centroidi sono generalmente equidistanti dal centroide costruito sull'intero dataset. Nelle variabili categoriche, si nota che la tendenza è di seguire il dato moda della variabile: tuttavia, nelle variabili in cui il valore moda è meno netto, e quindi ci sono diversi altri valori con frequenza simile, con l'aggiunta di cluster i centroidi diventano più rappresentativi della popolazione. Ad esempio, le variabili 'color' e 'interior' si comportano in senso opposto tra di loro: la prima presenta 'black' come moda, ma i valori 'blue', 'white', 'gray' e 'silver' sono sensibilmente più frequenti di altri valori; nel caso di interior, 'black' e 'grey' rappresentano il 75% delle istanze, e dunque i cluster non si separano dal valore di moda 'black'.

kMeans
======

Number of iterations: 13
Within cluster sum of squared errors: 1586324.145729521

Initial starting points (random):

Cluster 0: chevrolet,tahoe,2000,suv,automatic,onemain-rem/ramos-autos-&-trucks-inc,ga,2700,1.9,173327,brown,brown,2175
Cluster 1: nissan,pathfinder,2007,suv,automatic,santander-consumer,mo,5000,3.3,153081,blue,gray,5100
Cluster 2: toyota,camry-hybrid,2014,sedan,automatic,h&b-auto-center-inc,nj,15300,4.4,26135,gray,gray,17600

Missing values globally replaced with mean/mode

Final cluster centroids:

| | | Cluster# | | |
|---|---|---|---|---|
| Attribute | Full Data | 0 | 1 | 2 |
| | (307428.0) | (75548.0) | (88269.0) | (143611.0) |
| make | ford | chevrolet | ford | ford |
| model | altima | silverado-1500 | escape | altima |
| year | 2010.2173 | 2008.1371 | 2011.8099 | 2010.3327 |
| body | sedan | suv | suv | sedan |
| transmission | automatic | automatic | automatic | automatic |
| seller | nissan-infiniti-lt | the-hertz-corporation | ford-motor-credit-companyllc | nissan-infiniti-lt |
| state | fl | ga | fl | fl |
| sellingprice | 13698.6613 | 10029.1332 | 19537.6341 | 12040.1896 |
| condition | 3.4293 | 2.8288 | 4.0741 | 3.3488 |
| odometer | 66650.3988 | 96568.0542 | 45522.3758 | 63898.0361 |
| color | black | black | black | gray |
| interior | black | black | black | black |
| mmr | 13840.5897 | 10430.1871 | 19431.3983 | 12198.334 |

Figure 5: Three Clusters Model

kMeans
======

Number of iterations: 34
Within cluster sum of squared errors: 1537458.7299178573

Initial starting points (random):

Cluster 0: chevrolet,tahoe,2000,suv,automatic,onemain-rem/ramos-autos-&-trucks-inc,ga,2700,1.9,173327,brown,brown,2175
Cluster 1: nissan,pathfinder,2007,suv,automatic,santander-consumer,mo,5000,3.3,153081,blue,gray,5100
Cluster 2: toyota,camry-hybrid,2014,sedan,automatic,h&b-auto-center-inc,nj,15300,4.4,26135,gray,gray,17600
Cluster 3: chevrolet,sonic,2014,hatchback,automatic,avis-corporation,mo,10700,4.1,35226,burgundy,gray,9600

Missing values globally replaced with mean/mode

Final cluster centroids:

| | | Cluster# | | | |
|---|---|---|---|---|---|
| Attribute | Full Data | 0 | 1 | 2 | 3 |
| | (307428.0) | (58220.0) | (55037.0) | (102117.0) | (92054.0) |
| make | ford | chevrolet | ford | ford | chevrolet |
| model | altima | equinox | escape | fusion | impala |
| year | 2010.2173 | 2009.2806 | 2010.4358 | 2011.9023 | 2008.8099 |
| body | sedan | suv | suv | sedan | sedan |
| transmission | automatic | automatic | automatic | automatic | automatic |
| seller | nissan-infiniti-lt | the-hertz-corporation | ford-motor-credit-companyllc | nissan-infiniti-lt | avis-corporation |
| state | fl | ga | fl | fl | fl |
| sellingprice | 13698.6613 | 13801.8439 | 16506.0406 | 15791.6782 | 9633.1169 |
| condition | 3.4293 | 3.2862 | 3.7296 | 3.8554 | 2.8675 |
| odometer | 66650.3988 | 80649.4905 | 64074.5036 | 43600.0944 | 84906.7524 |
| color | black | black | blue | gray | white |
| interior | black | black | black | black | black |
| mmr | 13840.5897 | 13974.9077 | 16550.8145 | 15766.1927 | 9999.1551 |

Figure 6: Four Clusters Model

# 5 Conclusions

In conclusione, il presente caso di studio ha applicato nozioni di rappresentazione di conoscenza e modelli di apprendimento non supervisionato al fine di trarre preziose informazioni sul dataset di partenza. La knowledge base ha permesso di rappresentare in modo strutturato le caratteristiche e relazioni tra i diversi modelli di auto, facilitando l'accesso alla conoscenza specifica del dominio. Parallelamente, il clustering ha rivelato gruppi di dati, individuando modelli su diverse caratteristiche descritte dal dataset.

In futuro, gli approcci potrebbero essere utilizzati in coordinazione per comporre un più sofisticato recommender system per potenziali clienti, con query più indicate per il target e applicazioni di modelli di clustering più complessi al fine di riconoscere connessioni più significative.

# Notes

[1] Il dataset e la documentazione dell'autore possono essere reperiti al seguente indirizzo: https://www.kaggle.com/datasets/tunguz/used-car-auction-prices

[2] Più informazioni al seguente indirizzo: https://site.manheim.com/en/services/valuation.html