UNIVERSITÀ POLITECNICA DELLE MARCHE

FACOLTÀ DI ECONOMIA "GIORGIO FUÀ"

Corso di Laurea triennale in

Economia aziendale

REGRESSIONE LINEARE IN PYTHON: UN MODELLO PREVISIVO PER LA DETERMINAZIONE DEL PREZZO DELLE AUTO

LINEAR REGRESSION WITH PYTHON: A PREDICTIVE MODEL TO DETERMINE CAR PRICES

Relatore:
Prof.ssa Claudia Pigini

Rapporto Finale di:
Silvia Palanca

Anno Accademico 2023/2024

Indice

INDICE 3

INTRODUZIONE 4

I. DESCRIZIONE DEI METODI 6

II. DESCRIZIONE DEL DATASET 8

II.1 PANORAMICA 8

II.2 VARIABILI 9

III. APPLICAZIONE DEI METODI 20

III.1 CORRELAZIONE 20

III.2 OLS 22

III.3 VIF 23

III.4 MODELLO DI REGRESSIONE LINEARE 25

CONCLUSIONE 29

BIBLIOGRAFIA E SITOGRAFIA 31

INTRODUZIONE

Sebbene l'analisi predittiva esista ormai da decenni, è una tecnica più che mai diffusa: sempre più aziende stanno ricorrendo a modelli statistici per aumentare i loro profitti, ridurre i rischi e, in definitiva, acquisire vantaggio competitivo.

Implementare i propri strumenti con metodi di previsione è quindi una conditio sine qua non per sopravvivere nel contesto competitivo affollato in cui l'impresa è inserita e svolgere nel modo più efficiente ed efficace le proprie attività.

Un settore che fa ampio uso di questo strumento, essendo ad alto contenuto tecnologico e quindi particolarmente soggetto all'innovazione continua, è quello automobilistico. Le grandi aziende del settore devono infatti prevedere i prezzi e le vendite dei propri prodotti, già circolanti o da lanciare, sia nel breve che nel medio-lungo periodo. Tale processo permette ai manager di prendere decisioni migliori mettendo a frutto l'enorme mole di dati a loro disposizione, che altrimenti sarebbe indecifrabile.

In questo contesto si inseriscono i modelli di regressione lineare. Essi sono tra i più applicati essendo un metodo collaudato per generare previsioni di relativamente semplice interpretazione.

Questo elaborato approfondisce in particolare l'applicazione della regressione multipla nella previsione del prezzo di un'autovettura nuova per supportare un nuovo entrante cinese nel mercato americano. Prima di tutto andrà a descrivere la regressione lineare e le sue caratteristiche in generale, successivamente si concentrerà sull'applicazione di suddetto metodo su un dataset relativo ad auto nuove. Le fasi di stima e diagnostica, così come la

visualizzazione e lettura del dataset verranno svolte tramite l'utilizzo del software Python, ormai il più diffuso in questi ambiti dato che si presta particolarmente a tale tipo di analisi.

È indispensabile precisare che l'argomento delle pagine che seguono, così come appena descritto, è stato già ampliamente trattato da libri e articoli. Per citarne uno, Makoto Ohta e Zvi Griliches con il capitolo "Automobile Prices Revisited: Extensions of the Hedonic Hypothesis" nel libro "Household Production and Consumption" del 1976, hanno aperto la strada alla letteratura successiva.

Lo scopo delle pagine che seguiranno è quello di capire quali sono le variabili che descrivono il prezzo di un'auto, che relazione hanno con esso e se la regressione lineare è il metodo più preciso per prevederlo. Queste considerazioni permetteranno di capire le dinamiche di prezzo del mercato e le eventuali strategie da adottare da parte dell'azienda cinese che vuole introdurvisi.

DESCRIZIONE DEI METODI

Come precedentemente anticipato, il metodo applicato in questo elaborato sarà la regressione lineare. Quindi andremo a cercare quelle variabili che influenzano il prezzo dell'auto e cercheremo di capire la relazione che le lega ad esso.

Data la copiosa presenza di testi che spiegano nel dettaglio il metodo scelto, andremo a fornire una spiegazione succinta al riguardo rimandando al testo Basic Econometrics di Riccardo Lucchetti per ulteriori approfondimenti.

Prima di cominciare è doveroso ricordare che i modelli sono una rappresentazione matematica della realtà, citando l'econometrico George Box "All models are wrong, but some are useful". Quindi la reale relazione tra le variabili indipendenti e quella dipendente è un fenomeno inosservabile. La regressione multipla ci permette di determinare una funzione lineare che esprima nel modo migliore possibile il legame (in media) tra le variabili indipendenti - caratteristiche dell'auto - e la variabile dipendente - prezzo. Nel caso di *k* variabili indipendenti l'equazione è la seguente:

$$\boldsymbol{Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon_i}$$

Dove

- $\boldsymbol{Y}$ e $\boldsymbol{X_1}$, $\boldsymbol{X_2}$, …, $\boldsymbol{X_k}$ sono i valori, rispettivamente, della variabile dipendente e delle k variabili indipendenti rilevate con riferimento alla i-esima unità statistica.
- $\boldsymbol{\beta_0}$ è la constante o intercetta.
- $\boldsymbol{\beta_1}$, $\boldsymbol{\beta_2}$,…, $\boldsymbol{\beta_k}$ sono tutti i coefficienti di regressione parziale che indicano di quanto varia in media la $\boldsymbol{Y}$ quando $\boldsymbol{X_j}$ aumenta di un'unità, a parità di valori delle altre variabili esplicative.

- $\boldsymbol{\varepsilon_i}$ è il residuo non spiegato relativo all'osservazione i-esima. In altre parole, tutto ciò che le variabili indipendenti non riescono a spiegare finisce nell'errore che traduce l'incapacità del modello di riprodurre con esattezza la realtà osservata.

Geometricamente parlando, con questo modello stiamo cercando l'iperpiano - dato che usiamo più variabili - che descrive al meglio i dati osservati. Per farlo si segue un criterio tanto logico quanto efficace: minimizzare l'errore commesso. L'errore, sempre in termini geometrici, è la distanza tra tutti i punti realmente osservati e l'iperpiano che li sintetizza. Parliamo allora di metodo dei minimi quadrati perchè andiamo a minimizzare la somma dei quadrati di tutti gli errori di stima commessi. La formula che segue è di conseguenza il punto di partenza:

$$min \sum \varepsilon_i^2$$

Per questa ragione parliamo di OLS (Ordinary Least Squares).

Nell'applicare queste formule e costruire il modello, come precedentemente anticipato, ci faremo aiutare dal software Python che oltre a semplificare l'analisi la arricchisce con i suoi strumenti avanzati.

DESCRIZIONE DEL DATASET

Il dataset su cui si dipanerà la trattazione è relativo a numerose tipologie di auto presenti nel mercato americano e ne contiene varie caratteristiche così come il prezzo.

Di seguito una panoramica delle variabili, la grandezza del campione e i tipi di dati al suo interno. Tutte le figure che seguiranno sono output del software utilizzato nell'analisi.

II.1 PANORAMICA

Il totale delle osservazioni è 205 e le variabili indipendenti disponibili sono 25.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 205 entries, 0 to 204
Data columns (total 26 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   car_ID            205 non-null    int64
 1   symboling         205 non-null    int64
 2   CarName           205 non-null    object
 3   fueltype          205 non-null    object
 4   aspiration        205 non-null    object
 5   doornumber        205 non-null    object
 6   carbody           205 non-null    object
 7   drivewheel        205 non-null    object
 8   enginelocation    205 non-null    object
 9   wheelbase         205 non-null    float64
 10  carlength         205 non-null    float64
 11  carwidth          205 non-null    float64
 12  carheight         205 non-null    float64
 13  curbweight        205 non-null    int64
 14  enginetype        205 non-null    object
 15  cylindernumber    205 non-null    object
 16  enginesize        205 non-null    int64
 17  fuelsystem        205 non-null    object
 18  boreratio         205 non-null    float64
 19  stroke            205 non-null    float64
 20  compressionratio  205 non-null    float64
 21  horsepower        205 non-null    int64
 22  peakrpm           205 non-null    int64
 23  citympg           205 non-null    int64
 24  highwaympg        205 non-null    int64
 25  price             205 non-null    float64
dtypes: float64(8), int64(8), object(10)
memory usage: 41.8+ KB
```

Fig. II.1

II.2 VARIABILI

Prima di procedere con la costruzione del modello vero e proprio, andiamo ad analizzare le variabili una ad una.

Price

Il prezzo dell'auto è la variabile dipendente ed è calcolata in dollari, essendo il dataset relativo al mercato americano. Di seguito alcune statistiche descrittive e la distribuzione.

```
count      205.000000
mean     13276.710571
std       7988.852332
min       5118.000000
25%       7788.000000
50%      10295.000000
75%      16503.000000
max      45400.000000
Name: price, dtype: float64
```

Fig. II.2

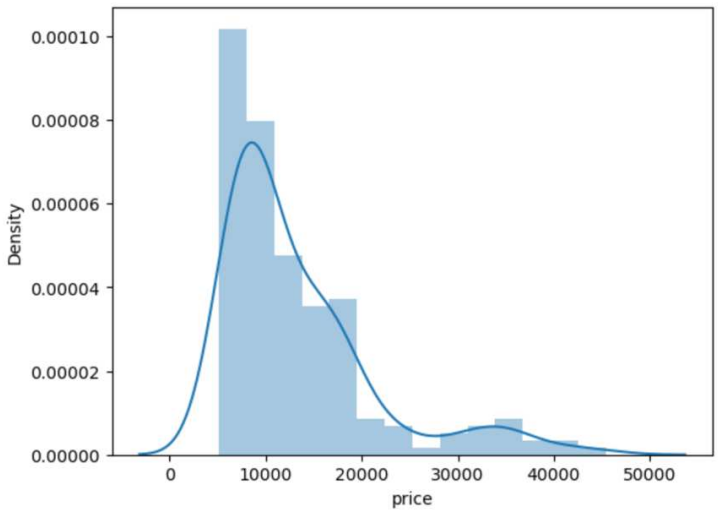

Fig. II.3

Symboling

La prima variabile esplicativa si riferisce a una valutazione del rischio assicurativo assegnata all'auto. Un valore di 3 significa che l'auto è rischiosa, mentre uno di -3 indica che probabilmente è sicura. Di seguito la distribuzione di frequenza e il prezzo medio per categoria.

```
symboling
 0    67
 1    54
 2    32
 3    27
-1    22
-2     3
Name: count, dtype: int64
```

Fig. II.4

```
symboling
-2    15781.666667
-1    17330.681818
 0    14366.965179
 1    10037.907407
 2    10109.281250
 3    17221.296296
Name: price, dtype: float64
```

Fig. II.5

Car Brand

Un altro importante fattore che determina il prezzo di un'auto è la casa produttrice e la sua popolarità. Perciò una seconda variabile indipendente è il marchio dell'auto. Visualizziamo un diagramma e il prezzo medio per marchio.

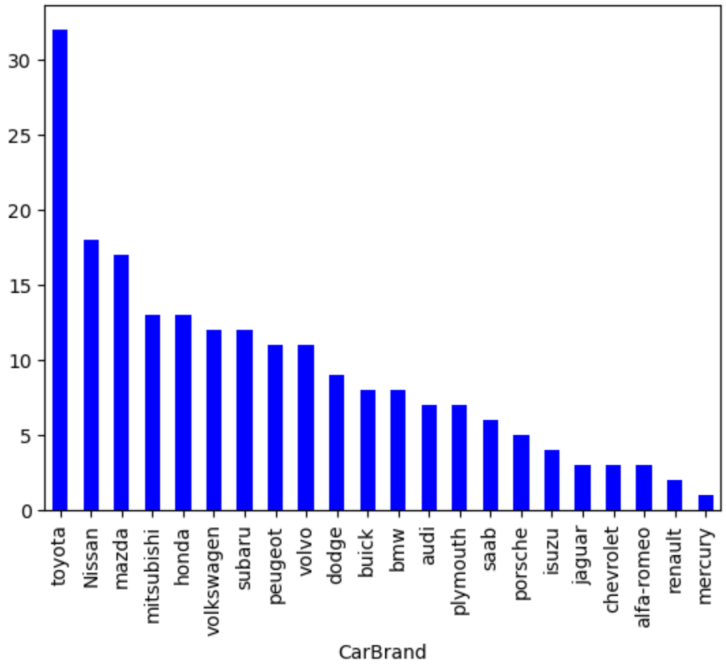

Fig. II.6

```
CarBrand
chevrolet       6007.000000
dodge           7875.444444
plymouth        7963.428571
honda           8184.692308
subaru          8541.250000
isuzu           8916.500000
mitsubishi      9239.769231
renault         9595.000000
toyota          9885.812500
volkswagen     10077.500000
Nissan         10415.666667
mazda          10652.882353
saab           15223.333333
peugeot        15489.090909
alfa-romeo     15498.333333
mercury        16503.000000
audi           17859.166714
volvo          18063.181818
bmw            26118.750000
porsche        31400.500000
buick          33647.000000
jaguar         34600.000000
Name: price, dtype: float64
```

Fig. II.7

Dato che si tratta di una variabile categorica, per utilizzare questa informazione creiamo delle dummy per ogni categoria.

Fuel Type

Nel dataset analizzato, il tipo di carburante delle auto si divide tra diesel e gas. Di seguito la distribuzione di frequenza e il prezzo medio per categoria.

```
fueltype
gas       185
diesel     20
Name: count, dtype: int64
```

Fig. II.8

```
fueltype
diesel    15838.1500
gas       12999.7982
Name: price, dtype: float64
```

Fig. II.9

Anche qui è necessario creare una dummy.

Aspiration

Questa variabile esplicativa fa riferimento al tipo di aspirazione utilizzata nell'autovettura. Le categorie possibili sono "standard" e "turbo".

```
aspiration
std      168
turbo     37
Name: count, dtype: int64
```

Fig. II.10

```
aspiration
std      12611.270833
turbo    16298.166676
Name: price, dtype: float64
```

Fig. II.11

Come per tutte le variabili categoriche, creiamo una dummy.

Number of Doors

Il numero di porte dell'auto è un'ulteriore informazione che ci permette di predirne il prezzo. Si tratta di una variabile dicotomica: due o quattro porte. Ne consegue che per utilizzarla nel modello c'è bisogno di creare una dummy.

```
doornumber
four    115
two      90
Name: count, dtype: int64
```

Fig. II.12

```
doornumber
four    13501.152174
two     12989.924078
Name: price, dtype: float64
```

Fig. II.13

Car Body

Con questa colonna si prende in considerazione la carrozzeria: parliamo di conseguenza di una variabile categorica che tratteremo come le precedenti.

```
carbody
sedan          96
hatchback      70
wagon          25
hardtop         8
convertible     6
Name: count, dtype: int64
```

Fig. II.14

```
carbody
hatchback      10376.652386
wagon          12371.960000
sedan          14344.270833
convertible    21890.500000
hardtop        22208.500000
Name: price, dtype: float64
```

Fig. II.15

Drive Wheel

Tra le auto comprese nel database ci possono essere tre diverse configurazioni per quanto concerne le ruote motrici: quattro ruote motrici, solo le anteriori o solo le posteriori. Generiamo quindi due dummy variables.

```
drivewheel
fwd    120
rwd     76
4wd      9
Name: count, dtype: int64
```

Fig. II.16

```
drivewheel
4wd    11087.463000
fwd     9239.308333
rwd    19910.809211
Name: price, dtype: float64
```

Fig. II.17

Engine Location

Il motore può essere posizionato davanti o dietro. Ecco che introduciamo un'ulteriore variabile categorica che richiede gli stessi passaggi delle altre.

```
enginelocation
front    202
rear       3
Name: count, dtype: int64
```

Fig. II.18

```
enginelocation
front    12961.097361
rear     34528.000000
Name: price, dtype: float64
```

Fig. II.19

Wheel Base

Passiamo ora a una variabile continua, la quale si riferisce all'interasse, cioè la distanza tra l'asse della ruota anteriore e quello della ruota posteriore. Possiamo allora prendere in esame alcune statistiche descritte e visualizzarla tramite un box plot.

```
count    205.000000
mean      98.756585
std        6.021776
min       86.600000
25%       94.500000
50%       97.000000
75%      102.400000
max      120.900000
Name: wheelbase, dtype: float64
```

Fig. II.20

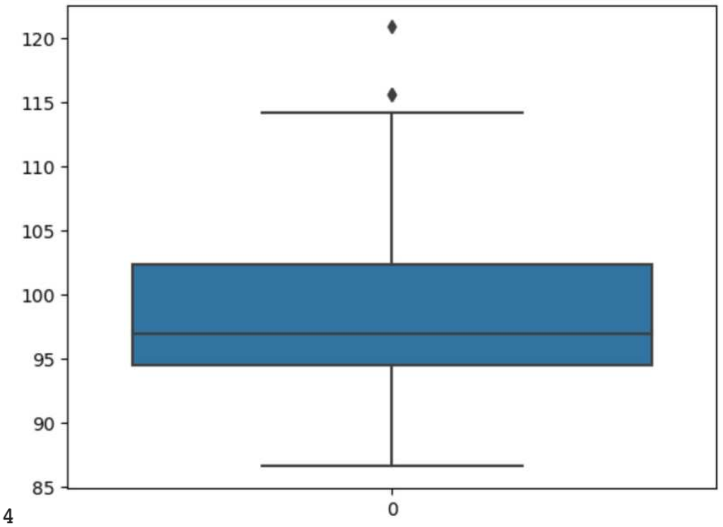

Fig. II.21

Car Length

Il prezzo finale è influenzato anche dalla lunghezza dell'auto. Come per le altre variabili continue, consideriamo le statistiche descrittive e una visualizzazione tramite box plot.

```
count    205.000000
mean     174.049268
std       12.337289
min      141.100000
25%      166.300000
50%      173.200000
75%      183.100000
max      208.100000
Name: carlength, dtype: float64
```

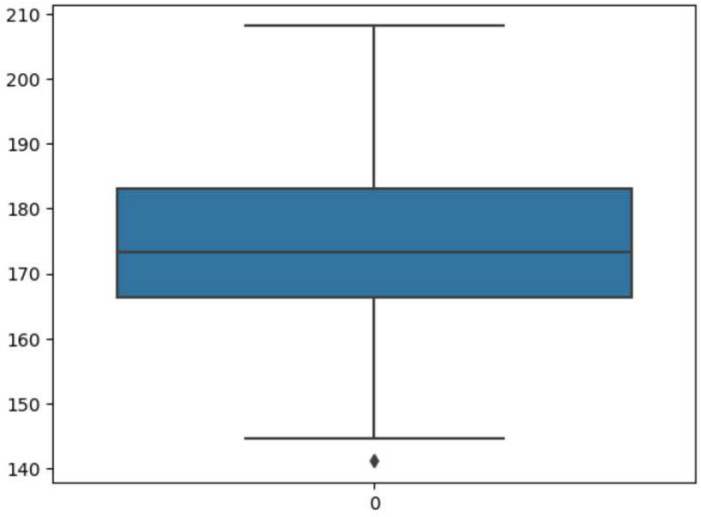

Fig. II.22 Fig. II.23

Car Width

Questa variabile indipendente e continua riguarda la profondità dell'autovettura.

```
count    205.000000
mean      65.907805
std        2.145204
min       60.300000
25%       64.100000
50%       65.500000
75%       66.900000
max       72.300000
Name: carwidth, dtype: float64
```

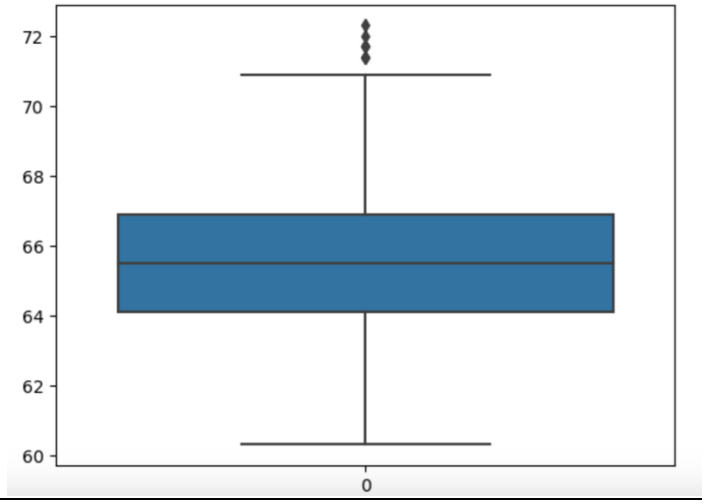

Fig. II.24 Fig. II.25

Car Height

Anche l'altezza della macchina è un'informazione compresa nel dataset essendo correlata con il prezzo finale.

```
count    205.000000
mean      53.724878
std        2.443522
min       47.800000
25%       52.000000
50%       54.100000
75%       55.500000
max       59.800000
Name: carheight, dtype: float64
```

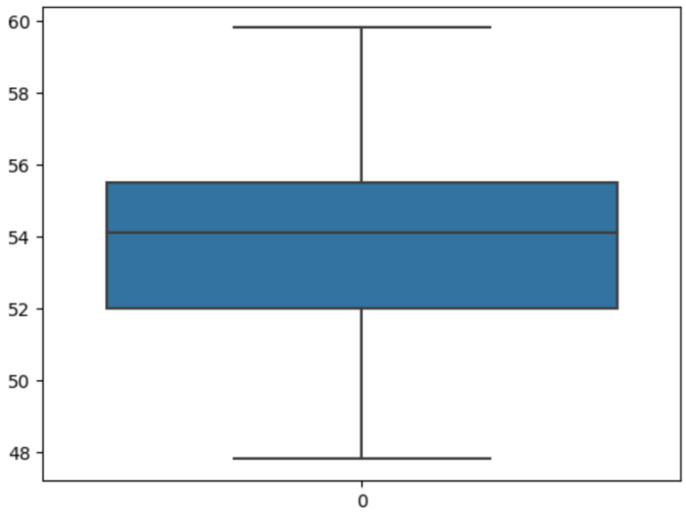

Fig. II.26 Fig. II.27

Car Weight

A completare le ultime variabili elencate, che si riferiscono tutte alle caratteristiche esteriori del veicolo, abbiamo il suo peso. Esso è considerato ovviamente escludendo passeggeri e bagagli.

```
count     205.000000
mean     2555.565854
std       520.680204
min      1488.000000
25%      2145.000000
50%      2414.000000
75%      2935.000000
max      4066.000000
Name: curbweight, dtype: float64
```

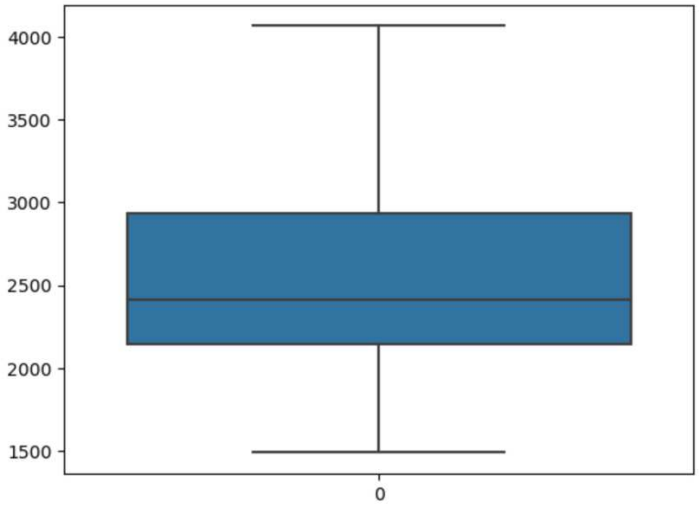

Fig. II.28 Fig. II.29

Engine Type

Passiamo ora a caratteristiche interne relative al motore. I vari tipi di motore possibili sono inclusi nel dataset comprendendo diverse categorie. Come al solito evidenziamo la tabella di frequenza, il prezzo medio per categoria e creiamo le dummy per ognuna.

```
enginetype
ohc      148
ohcf      15
ohcv      13
dohc      12
l         12
rotor      4
dohcv      1
Name: count, dtype: int64
```

Fig. II.30

```
enginetype
ohc      11574.048426
rotor    13020.000000
ohcf     13738.600000
l        14627.583333
dohc     18116.416667
ohcv     25098.384615
dohcv    31400.500000
Name: price, dtype: float64
```

Fig. II.31

Cylinder number

Un fattore fondamentale da tenere in considerazione quando si parla di motore è il numero di cilindri. Anche questa è una variabile categorica che trattiamo al pari delle altre.

```
cylindernumber
four      159
six        24
five       11
eight       5
two         4
three       1
twelve      1
Name: count, dtype: int64
```

Fig. II.32

```
cylindernumber
three     5151.000000
four     10285.754717
two      13020.000000
five     21630.469727
six      23671.833333
twelve   36000.000000
eight    37400.100000
Name: price, dtype: float64
```

Fig. II.33

Engine size

Un'altra caratteristica significativa, con riferimento al motore dell'autovettura, è la sua dimensione. Si tratta di una variabile indipendente e continua.

```
count    205.000000
mean     126.907317
std       41.642693
min       61.000000
25%       97.000000
50%      120.000000
75%      141.000000
max      326.000000
Name: enginesize, dtype: float64
```

Fig. II.34

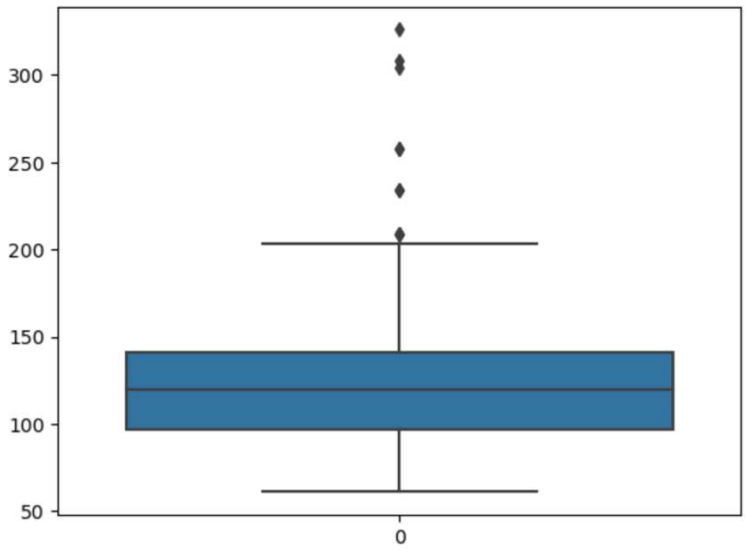

Fig. II.35

Fuel System

Per permettere al carburante di raggiungere il motore, ogni auto è dotata di un sistema di alimentazione. Ne esistono vari tipi e per questo la variabile esplicativa descritta di seguito è categorica, con tutto ciò che questo implica.

```
fuelsystem
mpfi    94
2bbl    66
idi     20
1bbl    11
spdi     9
4bbl     3
mfi      1
spfi     1
Name: count, dtype: int64
```

Fig. II.36

```
fuelsystem
2bbl     7478.151515
1bbl     7555.545455
spdi    10990.444444
spfi    11048.000000
4bbl    12145.000000
mfi     12964.000000
idi     15838.150000
mpfi    17754.602840
Name: price, dtype: float64
```

Fig. II.37

Bore Ratio

Un parametro motoristico tra i più importanti da considerare, insieme a quello che segue, è l'alesaggio, che indica il diametro interno del cilindro del motore. Trattandosi di una misura, le analisi di seguito riportate sono quelle tipiche di una variabile continua.

```
count    205.000000
mean       3.329756
std        0.270844
min        2.540000
25%        3.150000
50%        3.310000
75%        3.580000
max        3.940000
Name: boreratio, dtype: float64
```

Fig. II.38

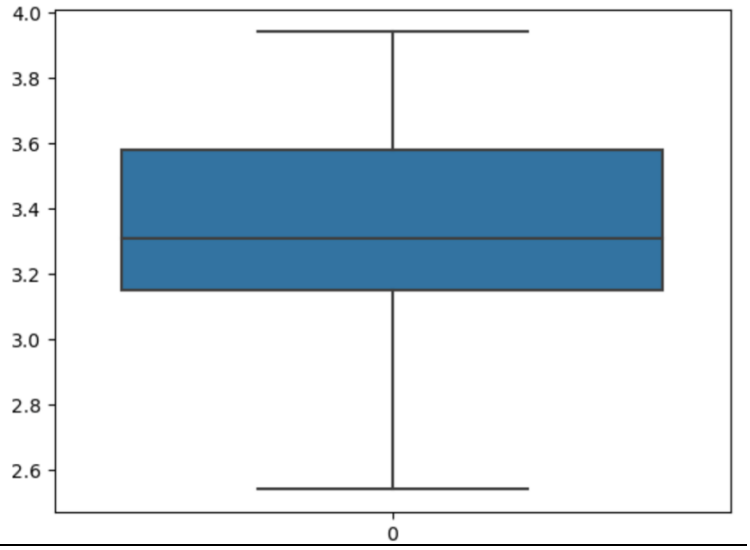

Fig. II.39

Stroke

Il secondo parametro motoristico indispensabile è la corsa, cioè la distanza tra l'estremità superiore e inferiore del pistone. Insieme all'alesaggio serve per calcolare la cilindrata del motore. Anche in questo caso parliamo di una misura.

```
count    205.000000
mean       3.255415
std        0.313597
min        2.070000
25%        3.110000
50%        3.290000
75%        3.410000
max        4.170000
Name: stroke, dtype: float64
```

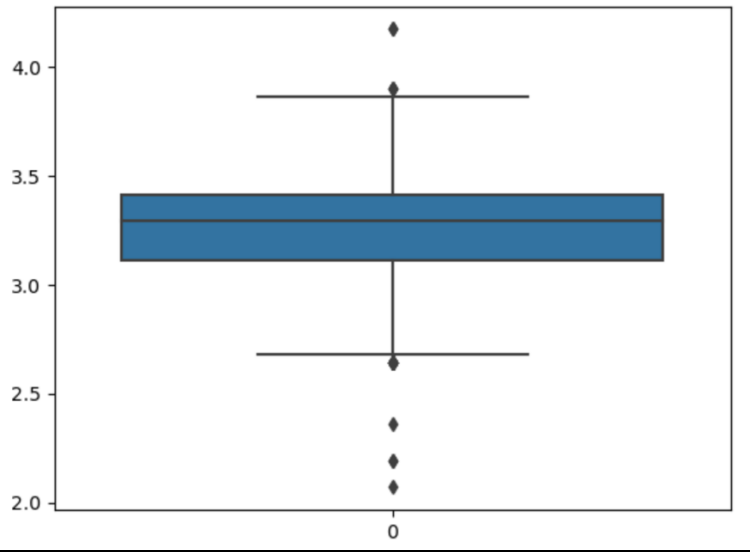

Fig. II.40 Fig. II.41

Compression Ratio

Il rapporto di compressione è un indice molto importante in quanto permette di valutare l'efficienza di un motore. Anche questa è una variabile numerica.

```
count    205.000000
mean      10.142537
std        3.972040
min        7.000000
25%        8.600000
50%        9.000000
75%        9.400000
max       23.000000
Name: compressionratio, dtype: float64
```

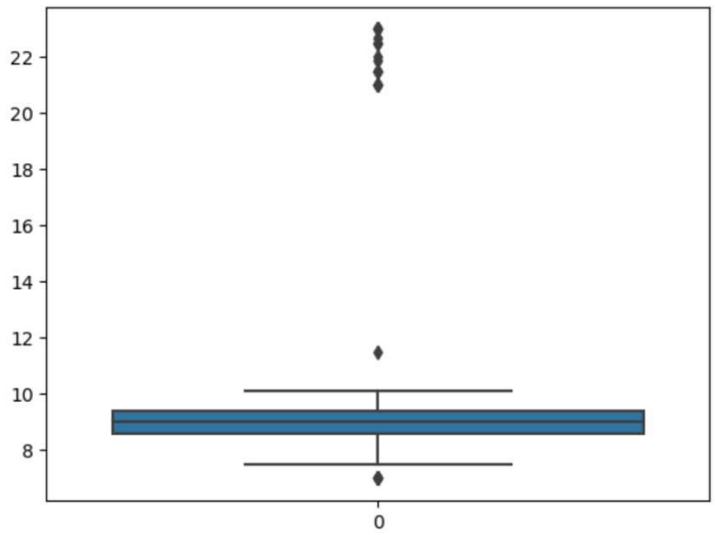

Fig. II.42 Fig. II.43

Horse-Power

L'unità di misura della potenza del motore è il cavallo vapore. Consideriamo quindi le statistiche e visualizzazioni di questa variabile continua.

```
count    205.000000
mean     104.117073
std       39.544167
min       48.000000
25%       70.000000
50%       95.000000
75%      116.000000
max      288.000000
Name: horsepower, dtype: float64
```

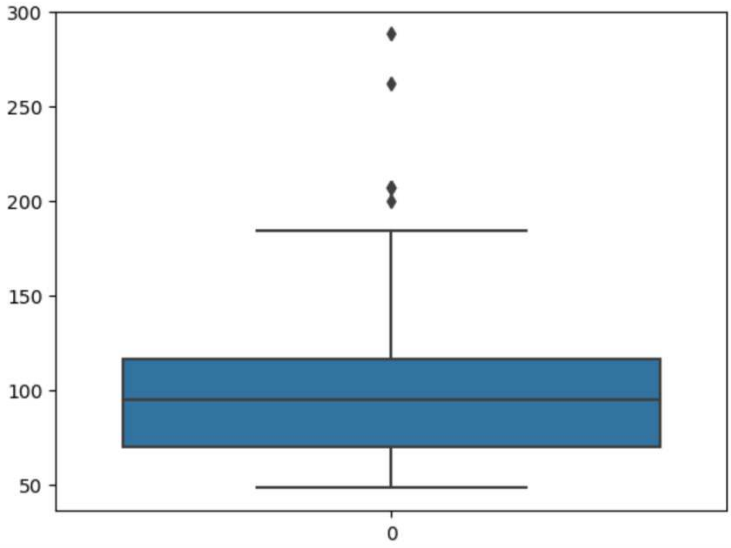

Fig. II.44 Fig. II.45

Peak RPM

Sempre parlando di motore, è necessario conoscere il numero di giri al minuto per avere un'idea della potenza da esso prodotta. Di seguito un'analisi di quest'ulteriore variabile numerica.

```
count     205.000000
mean     5125.121951
std       476.985643
min      4150.000000
25%      4800.000000
50%      5200.000000
75%      5500.000000
max      6600.000000
Name: peakrpm, dtype: float64
```

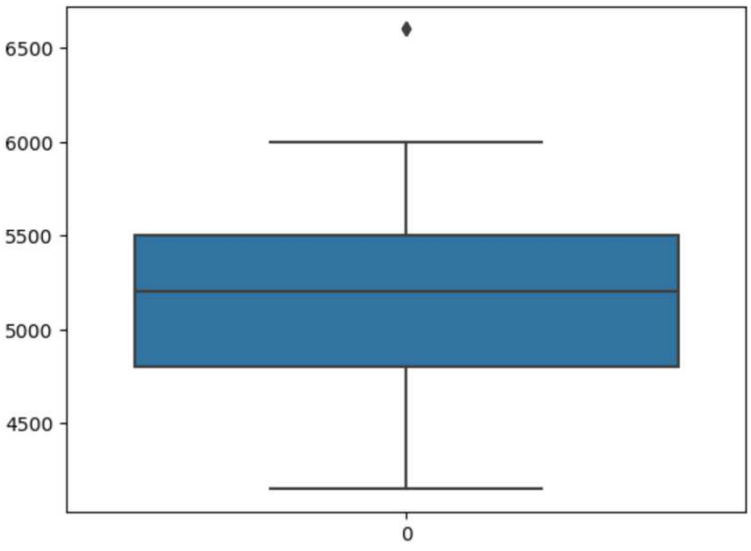

Fig. II.46 Fig. II.47

City Mileage

Arrivando al kilometraggio dell'auto, nel dataset viene fatta una distinzione tra miglia percorsi in città e in autostrada. L'unità di misura di questa variabile numerica è il miglio, essendo in riferimento al mercato americano.

```
count    205.000000
mean      25.219512
std        6.542142
min       13.000000
25%       19.000000
50%       24.000000
75%       30.000000
max       49.000000
Name: citympg, dtype: float64
```

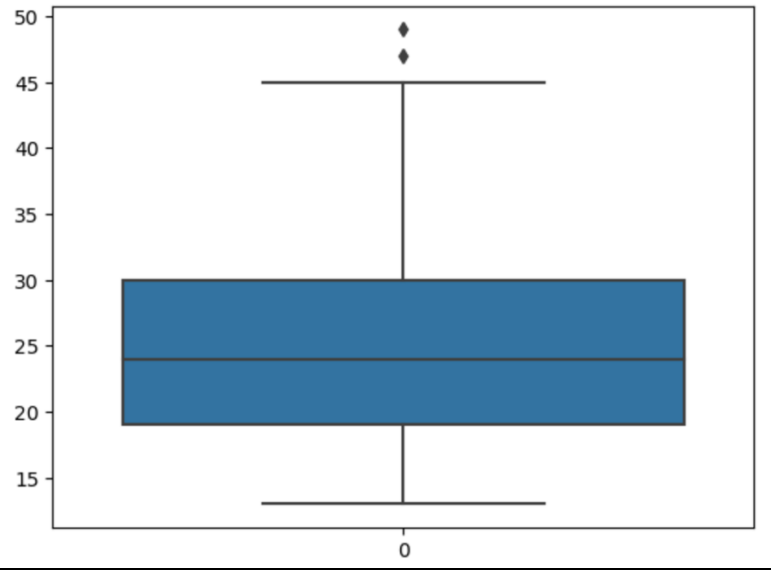

Fig. II.48 Fig. II.49

Highway Mileage

Passiamo infine alle miglia percorse in autostrada.

```
count    205.000000
mean      30.751220
std        6.886443
min       16.000000
25%       25.000000
50%       30.000000
75%       34.000000
max       54.000000
Name: highwaympg, dtype: float64
```

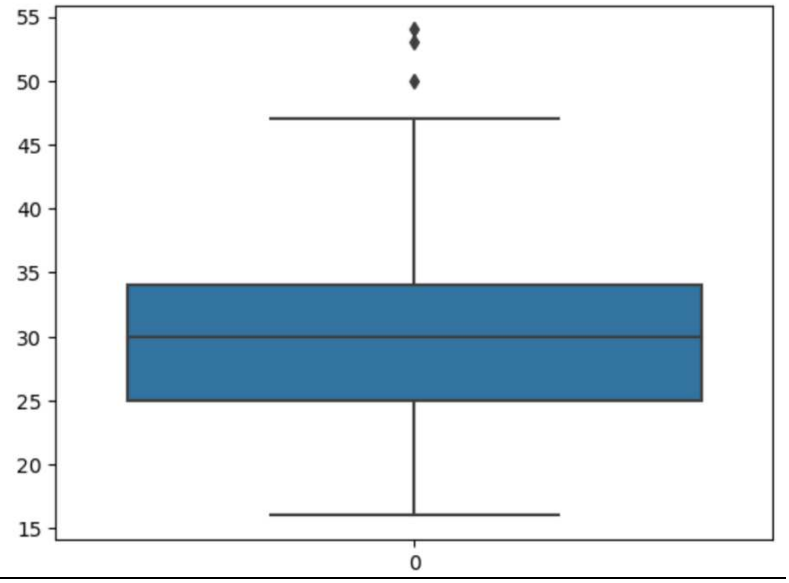

Fig. II.50 Fig. II.51

Con questo paragrafo si conclude la panoramica sulle variabili disponibili. Procediamo ora all'applicazione della regressione lineare sul dataset appena descritto.

APPLICAZIONE DEI METODI

III.1 CORRELAZIONE

Uno dei problemi principali nell'applicazione della regressione lineare è il rischio di correlazione tra le variabili esplicative. La correlazione è una misura che indica la relazione lineare tra due variabili casuali. È compresa tra -1 e 1, dove:

- -1 indica una relazione lineare inversa
- 0 significa che non è possibile stabilire tra le due variabili un andamento lineare
- 1 segnala una relazione lineare diretta.

Ignorare questa problematica significherebbe includere colonne ridondanti ottenibili come correlazione lineare delle altre. In altre parole, il modello sarebbe mispecificato.

Come è evidente dal capitolo precedente, le variabili a disposizione sono molte; si mostra necessario quindi, prima di procedere nell'applicazione della regressione lineare, escludere che alcune di esse siano correlate tra loro.

A questo scopo ci avvaliamo di un metodo semplice ma preciso: la matrice di correlazione. Ovviamente includerà solo i parametri numerici, ma è un ottimo strumento statistico per mostrare la forza e la direzione della correlazione tra variabili. Grazie al software utilizzato, possiamo rendere la matrice ancora più intuitiva, visualizzandola sotto forma di heatmap.

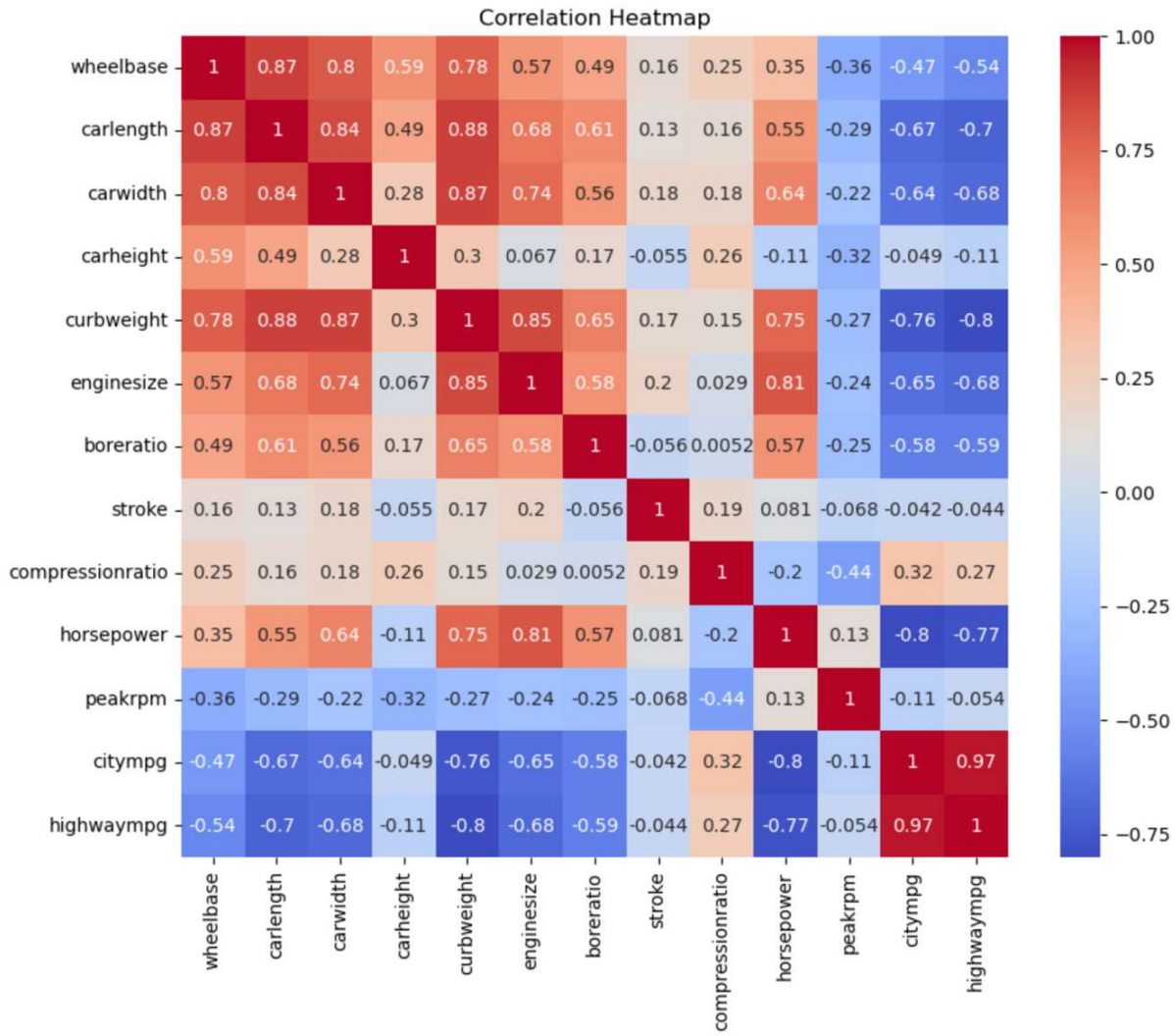

Fig. III.1

Quanto si evince dalla figura III.1 è purtroppo la presenza di alcune variabili correlate tra loro. Come era intuibile, le variabili che fanno riferimento alle caratteristiche esteriori dell'auto come altezza, profondità, lunghezza e peso sono correlate, così come quelle relative al kilometraggio in città e in autostrada. A seguito di un'attenta analisi, teniamo solo una variabile per ciascuna categoria: "wheel base" (interasse) per le dimensioni e "citympg" per il kilometraggio.

A questo punto possiamo procedere con il calcolo dei coefficienti del nostro modello di regressione lineare.

III.2 OLS

Utilizzando le variabili rimaste si possono calcolare i coefficienti del modello OLS, spiegato nel primo capitolo, con l'aiuto del software. Visualizziamo di seguito l'output di stima.

OLS Regression Results

Dep. Variable:	price	**R-squared:**	0.958
Model:	OLS	**Adj. R-squared:**	0.943
Method:	Least Squares	**F-statistic:**	62.21
Date:	Mon, 22 Jul 2024	**Prob (F-statistic):**	8.65e-80
Time:	10:44:17	**Log-Likelihood:**	-1806.9
No. Observations:	205	**AIC:**	3726.
Df Residuals:	149	**BIC:**	3912.
Df Model:	55		
Covariance Type:	nonrobust		

Fig. III.2

Omnibus:	55.411	**Durbin-Watson:**	1.695
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	211.813
Skew:	1.017	**Prob(JB):**	1.01e-46
Kurtosis:	7.546	**Cond. No.**	1.01e+16

Fig. III.3

	coef	std err	t	P>|t|	[0.025	0.975]
const	-1.171e+04	7771.524	-1.507	0.134	-2.71e+04	3641.717
symboling	-60.3733	270.811	-0.223	0.824	-595.500	474.753
wheelbase	218.2916	74.657	2.924	0.004	70.768	365.815
enginesize	101.2704	26.890	3.766	0.000	48.136	154.405
boreratio	-1759.6879	1957.734	-0.899	0.370	-5628.196	2108.821
stroke	-2134.2945	1052.042	-2.029	0.044	-4213.144	-55.445
compressionratio	-269.1651	515.729	-0.522	0.603	-1288.252	749.922
horsepower	44.6350	24.179	1.846	0.067	-3.144	92.414
peakrpm	2.0694	0.693	2.985	0.003	0.699	3.439
citympg	93.0802	70.300	1.324	0.188	-45.834	231.994
CarBrand_alfa-romeo	3571.9609	1808.595	1.975	0.050	-1.846	7145.768
CarBrand_audi	4434.2092	1801.897	2.461	0.015	873.638	7994.781
CarBrand_bmw	7410.5240	1522.465	4.867	0.000	4402.112	1.04e+04
CarBrand_buick	7453.6947	2151.103	3.465	0.001	3203.087	1.17e+04
CarBrand_chevrolet	-1341.1530	1548.138	-0.866	0.388	-4400.294	1717.989
CarBrand_dodge	-999.9020	940.336	-1.063	0.289	-2858.019	858.215
CarBrand_honda	2237.2410	1576.893	1.419	0.158	-878.721	5353.202
CarBrand_isuzu	290.8749	1303.063	0.223	0.824	-2283.994	2865.744
CarBrand_jaguar	7662.8869	2430.467	3.153	0.002	2860.252	1.25e+04
CarBrand_mazda	1308.1634	832.116	1.572	0.118	-336.109	2952.436
CarBrand_mercury	-1332.8604	2475.409	-0.538	0.591	-6224.300	3558.580
CarBrand_mitsubishi	-955.1562	946.607	-1.009	0.315	-2825.663	915.351
CarBrand_peugeot	1059.1363	1926.401	0.550	0.583	-2747.457	4865.729
CarBrand_plymouth	-1249.4094	1000.563	-1.249	0.214	-3226.535	727.716
CarBrand_porsche	8181.5775	2430.129	3.367	0.001	3379.610	1.3e+04
CarBrand_renault	1395.5535	1711.300	0.815	0.416	-1985.999	4777.106
CarBrand_saab	3610.5031	1492.321	2.419	0.017	661.657	6559.349
CarBrand_subaru	-31.6662	1336.602	-0.024	0.981	-2672.809	2609.476
CarBrand_toyota	94.5361	789.746	0.120	0.905	-1466.013	1655.085
CarBrand_volkswagen	1050.2754	1019.465	1.030	0.305	-964.202	3064.753

Fig. III. 5

CarBrand_volvo	1204.5815	1392.529	0.865	0.388	-1547.075	3956.238
fueltype_gas	-9443.9663	3842.606	-2.458	0.015	-1.7e+04	-1850.927
aspiration_turbo	2145.0924	863.705	2.484	0.014	438.399	3851.785
doornumber_two	-528.7043	518.119	-1.020	0.309	-1552.515	495.106
carbody_hardtop	-4031.5218	1213.335	-3.323	0.001	-6429.087	-1633.956
carbody_hatchback	-4141.0252	1162.888	-3.561	0.000	-6438.906	-1843.144
carbody_sedan	-4230.5132	1224.171	-3.456	0.001	-6649.490	-1811.536
carbody_wagon	-4513.0173	1318.112	-3.424	0.001	-7117.624	-1908.411
drivewheel_fwd	-788.4417	841.674	-0.937	0.350	-2451.600	874.717
drivewheel_rwd	25.5668	1148.182	0.022	0.982	-2243.257	2294.390
enginelocation_rear	2414.6672	1808.509	1.335	0.184	-1158.970	5988.305
enginetype_dohcv	-5232.2153	4990.479	-1.048	0.296	-1.51e+04	4629.038
enginetype_l	1501.0382	1590.572	0.944	0.347	-1641.954	4644.030
enginetype_ohc	2727.4603	1259.311	2.166	0.032	239.045	5215.875
enginetype_ohcf	2383.0010	1099.023	2.168	0.032	211.317	4554.685
enginetype_ohcv	-892.3399	1299.716	-0.687	0.493	-3460.596	1675.916
enginetype_rotor	2309.4339	2437.781	0.947	0.345	-2507.653	7126.520
cylindernumber_five	-4987.8666	3041.734	-1.640	0.103	-1.1e+04	1022.640
cylindernumber_four	-4435.3504	3734.811	-1.188	0.237	-1.18e+04	2944.685
cylindernumber_six	-3010.4461	2849.029	-1.057	0.292	-8640.164	2619.272
cylindernumber_three	441.9019	2993.903	0.148	0.883	-5474.089	6357.893
cylindernumber_twelve	-1.04e+04	5365.665	-1.938	0.055	-2.1e+04	205.679
cylindernumber_two	2309.4339	2437.781	0.947	0.345	-2507.653	7126.520
fuelsystem_2bbl	2886.1058	1596.879	1.807	0.073	-269.348	6041.560
fuelsystem_4bbl	1140.2553	2727.186	0.418	0.676	-4248.700	6529.211
fuelsystem_idi	-2270.9517	6386.815	-0.356	0.723	-1.49e+04	1.03e+04
fuelsystem_mfi	1547.7842	2837.715	0.545	0.586	-4059.578	7155.147
fuelsystem_mpfi	2203.0410	1679.637	1.312	0.192	-1115.944	5522.026
fuelsystem_spdi	1589.3795	1965.103	0.809	0.420	-2293.690	5472.449
fuelsystem_spfi	4189.4582	2851.134	1.469	0.144	-1444.421	9823.337

Fig. III.5

Guardando al valore di R^2 e dei test disponibili come Durbin-Watson e Jarque-Bera, il modello sembrerebbe avere risultati accettabili. Memori dei problemi di multicollinearità di cui sopra, è importante monitorare anche il valore dei VIF (variance inflation factor).

III.3 VIF

Il VIF è definito come:

$$VIF_j = \left(1 - R_j^2\right)^{-1}$$

R_j^2 è l'indice di determinazione calcolato sul modello lineare:

$$x_j = \sum_{i \neq j} x_i b_i + u_i$$

Cioè una regressione lineare di una colonna rispetto a tutte le altre. Ci sono problemi di multicollinearità qualora VIF_j sia > 20.

Osserviamo ora questo indicatore per le variabili del nostro modello.

	feature	VIF			
0	const	0.000000	30	CarBrand_volvo	5.536070
1	symboling	6.363093	31	fueltype_gas	inf
2	wheelbase	11.307710	32	aspiration_turbo	6.203554
3	enginesize	70.150365	33	doornumber_two	3.717047
4	boreratio	15.729950	34	carbody_hardtop	3.103958
5	stroke	6.089654	35	carbody_hatchback	17.096438
6	compressionratio	234.775564	36	carbody_sedan	20.978767
7	horsepower	51.149340	37	carbody_wagon	10.459611
8	peakrpm	6.118900	38	drivewheel_fwd	9.666874
9	citympg	11.834045	39	drivewheel_rwd	17.291132
10	CarBrand_alfa-romeo	2.651877	40	enginelocation_rear	inf
11	CarBrand_audi	6.020344	41	enginetype_dohcv	6.796934
12	CarBrand_bmw	4.887080	42	enginetype_l	inf
13	CarBrand_buick	9.756117	43	enginetype_ohc	17.897852
14	CarBrand_chevrolet	1.943079	44	enginetype_ohcf	inf
15	CarBrand_dodge	2.086716	45	enginetype_ohcv	5.640790
16	CarBrand_honda	8.303237	46	enginetype_rotor	inf
17	CarBrand_isuzu	1.826354	47	cylindernumber_five	26.414067
18	CarBrand_jaguar	4.789063	48	cylindernumber_four	136.486747
19	CarBrand_mazda	2.960554	49	cylindernumber_six	47.171727
20	CarBrand_mercury	1.672331	50	cylindernumber_three	inf
21	CarBrand_mitsubishi	2.992141	51	cylindernumber_twelve	7.857342
22	CarBrand_peugeot	inf	52	cylindernumber_two	inf
23	CarBrand_plymouth	1.856310	53	fuelsystem_2bbl	31.296887
24	CarBrand_porsche	7.900547	54	fuelsystem_4bbl	6.029766
25	CarBrand_renault	1.590658	55	fuelsystem_idi	inf
26	CarBrand_saab	3.557356	56	fuelsystem_mfi	2.197689
27	CarBrand_subaru	inf	57	fuelsystem_mpfi	39.380408
28	CarBrand_toyota	4.619228	58	fuelsystem_spdi	9.113129
29	CarBrand_volkswagen	3.220193	59	fuelsystem_spfi	2.218523

Fig. III.6

I valori osservati sono decisamente troppo alti. Procediamo allora nel rimuovere dal modello le variabili che causano problemi di collinearità.

III.4 MODELLO DI REGRESSIONE LINEARE

Il software può ora ricalcolare i coefficienti del modello che a questo punto non soffre più di mispecificazione.

OLS Regression Results

Dep. Variable:	price	**R-squared:**	0.946
Model:	OLS	**Adj. R-squared:**	0.931
Method:	Least Squares	**F-statistic:**	62.26
Date:	Mon, 22 Jul 2024	**Prob (F-statistic):**	1.66e-80
Time:	10:50:23	**Log-Likelihood:**	-1832.7
No. Observations:	205	**AIC:**	3757.
Df Residuals:	159	**BIC:**	3910.
Df Model:	45		
Covariance Type:	nonrobust		

Fig. III.7

Omnibus:	51.485	**Durbin-Watson:**	1.505
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	178.510
Skew:	0.971	**Prob(JB):**	1.73e-39
Kurtosis:	7.138	**Cond. No.**	2.51e+05

Fig. III.8

	coef	std err	t	P>|t|	[0.025	0.975]
const	9420.9320	6860.186	1.373	0.172	-4127.910	2.3e+04
symboling	-306.5146	257.387	-1.191	0.235	-814.853	201.824
enginesize	141.8698	14.943	9.494	0.000	112.358	171.381
boreratio	-4380.2594	1407.176	-3.113	0.002	-7159.426	-1601.093
stroke	-878.7422	959.019	-0.916	0.361	-2772.801	1015.317
peakrpm	1.4453	0.558	2.589	0.011	0.343	2.548
citympg	-46.8208	65.893	-0.711	0.478	-176.959	83.317
CarBrand_alfa-romeo	-383.0194	1574.897	-0.243	0.808	-3493.435	2727.396
CarBrand_audi	5517.2190	1739.476	3.172	0.002	2081.760	8952.678
CarBrand_bmw	9352.4482	1447.687	6.460	0.000	6493.271	1.22e+04
CarBrand_buick	8593.1254	1785.107	4.814	0.000	5067.546	1.21e+04
CarBrand_chevrolet	517.9075	1470.799	0.352	0.725	-2386.916	3422.731
CarBrand_dodge	-1312.6122	990.339	-1.325	0.187	-3268.529	643.304
CarBrand_honda	-379.2154	1193.808	-0.318	0.751	-2736.981	1978.550
CarBrand_isuzu	351.1994	1390.166	0.253	0.801	-2394.374	3096.772
CarBrand_jaguar	4614.1016	2094.312	2.203	0.029	477.843	8750.360
CarBrand_mazda	1846.0982	769.028	2.401	0.018	327.271	3364.925
CarBrand_mercury	1745.1072	2402.827	0.726	0.469	-3000.467	6490.681
CarBrand_mitsubishi	-852.0625	1019.426	-0.836	0.405	-2865.425	1161.300
CarBrand_plymouth	-1458.7447	1058.690	-1.378	0.170	-3549.654	632.165
CarBrand_porsche	1.159e+04	1617.691	7.165	0.000	8395.524	1.48e+04
CarBrand_renault	-581.7907	1772.555	-0.328	0.743	-4082.579	2918.998
CarBrand_saab	3839.1515	1324.649	2.898	0.004	1222.975	6455.328

Fig. III.9

CarBrand_toyota	-426.1329	666.763	-0.639	0.524	-1742.987	890.721
CarBrand_volkswagen	1078.1397	1045.933	1.031	0.304	-987.575	3143.854
CarBrand_volvo	2837.2466	1268.351	2.237	0.027	332.259	5342.234
aspiration_turbo	2930.3121	630.330	4.649	0.000	1685.413	4175.211
doornumber_two	-692.1958	555.089	-1.247	0.214	-1788.494	404.103
carbody_hardtop	-3135.6044	1266.466	-2.476	0.014	-5636.869	-634.340
carbody_hatchback	-3380.4217	1149.944	-2.940	0.004	-5651.557	-1109.286
carbody_sedan	-3438.0489	1236.529	-2.780	0.006	-5880.189	-995.909
carbody_wagon	-3629.9269	1325.400	-2.739	0.007	-6247.586	-1012.268
drivewheel_fwd	-212.3839	858.931	-0.247	0.805	-1908.769	1484.001
drivewheel_rwd	1198.3078	986.352	1.215	0.226	-749.735	3146.350
enginetype_dohcv	-2753.3780	2645.163	-1.041	0.299	-7977.566	2470.810
enginetype_ohc	-1312.4519	816.272	-1.608	0.110	-2924.585	299.682
enginetype_ohcv	-2892.2036	1157.427	-2.499	0.013	-5178.118	-606.289
cylindernumber_five	-1316.3380	1380.821	-0.953	0.342	-4043.455	1410.779
cylindernumber_six	-1096.1590	1148.318	-0.955	0.341	-3364.083	1171.765
cylindernumber_twelve	-7147.1303	3672.344	-1.946	0.053	-1.44e+04	105.735
fuelsystem_2bbl	-7.9601	908.569	-0.009	0.993	-1802.381	1786.461
fuelsystem_4bbl	4311.4314	2063.345	2.090	0.038	236.334	8386.529
fuelsystem_mfi	-830.3883	2616.576	-0.317	0.751	-5998.117	4337.340
fuelsystem_mpfi	0.1157	925.753	0.000	1.000	-1828.242	1828.474
fuelsystem_spdi	-1176.6672	1432.211	-0.822	0.413	-4005.279	1651.944
fuelsystem_spfi	952.7903	2732.223	0.349	0.728	-4443.341	6348.921

Fig. III.10

	feature	VIF			
0	const	2193.974061			
1	symboling	4.766092			
2	enginesize	17.962260			
3	boreratio	6.738606			
4	stroke	4.195988			
5	peakrpm	3.289824			
6	citympg	8.620874			
7	CarBrand_alfa-romeo	1.667358			
8	CarBrand_audi	4.652134	27	doornumber_two	3.537671
9	CarBrand_bmw	3.664020	28	carbody_hardtop	2.804110
10	CarBrand_buick	5.571046	29	carbody_hatchback	13.862365
11	CarBrand_chevrolet	1.454225	30	carbody_sedan	17.748365
12	CarBrand_dodge	1.919193	31	carbody_wagon	8.769162
13	CarBrand_honda	3.946075	32	drivewheel_fwd	8.347733
14	CarBrand_isuzu	1.723620	33	drivewheel_rwd	10.580824
15	CarBrand_jaguar	2.948543	34	enginetype_dohcv	1.583387
16	CarBrand_mazda	2.096731	35	enginetype_ohc	6.235309
17	CarBrand_mercury	1.306553	36	enginetype_ohcv	3.709231
18	CarBrand_mitsubishi	2.877452	37	cylindernumber_five	4.513582
19	CarBrand_plymouth	1.723269	38	cylindernumber_six	6.354279
20	CarBrand_porsche	2.902975	39	cylindernumber_twelve	3.051889
21	CarBrand_renault	1.415068	40	fuelsystem_2bbl	8.400918
22	CarBrand_saab	2.324116	41	fuelsystem_4bbl	2.861990
23	CarBrand_toyota	2.730176	42	fuelsystem_mfi	1.549348
24	CarBrand_volkswagen	2.810597	43	fuelsystem_mpfi	9.919568
25	CarBrand_volvo	3.808246	44	fuelsystem_spdi	4.013883
26	aspiration_turbo	2.739674	45	fuelsystem_spfi	1.689330

Fig. III.11

Questo modello ha dei valori ottimi sia in termini di VIF che per quanto concerne gli altri test. È quindi questa la funzione che permetterà all'azienda cinese di stimare i prezzi delle auto nel mercato americano in cui vuole entrare.

L'entrante cinese sa che il prezzo dipende positivamente da: dimensioni e picco dei giri del motore, il brand se famoso, l'aspirazione se turbo, la presenza di ruote motrici posteriori, il sistema di alimentazione con iniezione multiporta.

Dall'altro lato, hanno un'influenza negativa sul prezzo: un indicatore di rischio assicurativo alto, il rapporto di alesaggio, la corsa, il kilometraggio in città, il brand se non attraente, la

presenza di sole 2 porte, alcuni tipi di carrozzeria, la presenza di ruote motrici anteriori, alcuni tipi di motore e ammontare di cilindri, alcuni tipi di sistema di alimentazione.

CONCLUSIONE

Questo studio si è posto l'obiettivo di identificare e analizzare le variabili che descrivono il prezzo di un'autovettura applicando il metodo della regressione lineare multipla.

Nonostante le criticità emerse relativamente alla multicollinearità, la ricerca è riuscita, tramite l'utilizzo del software, a creare un algoritmo in grado di prevedere con accuratezza il prezzo basandosi sulle caratteristiche che sono emerse essere le più significative.

La regressione lineare è risultata uno strumento adeguato a capire le dinamiche del mercato e quindi un ottimo punto di partenza per elaborare delle tattiche efficaci da parte dell'azienda cinese, oggetto del nostro esame.

È doveroso precisare che ci sono molti altri modelli statistici, anche più complessi, che potrebbero essere utilizzati per analisi simili, come *K-nearest Neighbors*, *Decision Trees* e *Support Vector Machines*.

Ad ogni modo, il lavoro svolto ha confermato come i modelli predittivi siano ormai uno strumento non solo utile, ma necessario, per tutte le aziende che operano nel complicato mercato odierno.

BIBLIOGRAFIA E SITOGRAFIA

- Ohta, M., & Griliches, Z. (1976). Automobile prices revisited: Extensions of the hedonic hypothesis. In *Household production and consumption* (pp. 325-398). NBER.
- Lucchetti, R. J. (2019). *Basic econometrics*.
- Caroli, M. (2021). *Economia e gestione sostenibile delle imprese*. Mc Graw Hill.
- Palomba, G. (2015). *Elementi di statistica per l'econometria*. Clua Edizioni Ancona.
- Schlimmer, J. (1985). Automobile [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C5B01C.
- Università degli studi di Bari Aldo Moro [Online]: https://www.uniba.it.
- SAS Institute [Online]: https://www.sas.com.
- IBM [Online]: https://www.ibm.com.
- Università degli studi di Verona [Online]: https://www.dsu.univr.it.