***ANALIZA ȘI PREDICȚIA A PREȚURILOR AUTOMOBILELOR*** Dubciuc Damian

Universitatea Tehnică a Moldovei, Facultatea Calculatoare, Informatică și Microelectronică, Departamentul Informatica și Ingineria Sistemelor

**ABSTRACT**

În cadrul acestui studiu, s-a abordat analiza și predicția prețurilor automobilelor, utilizând setul de date "Used Cars Dataset" disponibil pe Kaggle. Scopul a fost optimizarea prețului de vânzare, stabilirea prețului de listă, analiza pieței și dezvoltarea de modele lineare pentru a estima probabilitatea de creștere sau descreștere a prețurilor. Inițierea analizei a presupus explorarea setului de date, factorizarea variabilelor categorice și realizarea unei analize descriptive.

Relațiile dintre variabile au fost evidențiate, iar modele semnificative care pot influența variația prețurilor au fost identificate. S-au efectuat, de asemenea, vizualizări grafice, inclusiv o diagramă cu bare empilate, unde prețurile automobilelor au fost reprezentate în funcție de vârsta lor. Modelarea lineară a constituit o componentă crucială a studiului, implicând dezvoltarea mai multor modele și explorarea diverselor combinații de variabile, precum vârsta, sexul, tipul de durere și nivelurile de colesterol.

Rezultatele au demonstrat că anumite variabile, cum ar fi vârsta și starea, sunt semnificativ asociate cu probabilitatea variației prețurilor. Modelele lineare au furnizat o înțelegere mai profundă a interacțiunilor dintre aceste variabile, susținând procesul de predicție. În concluzie, această analiză detaliată a setului de date "Used Cars Dataset" a contribuit la înțelegerea unor factori cheie. Studiul oferă perspective valoroase cu privire la modul în care diferite caracteristici ale mașinilor pot influența strategiile de marketing sau analiza pieței.

# INTRODUCERE

În contextul industriei automobilelor, analiza și predicția prețurilor devin componente esențiale pentru strategiile de piață și luarea deciziilor eficiente. O astfel de abordare se bazează pe utilizarea tehnicilor de analiză a datelor și modelelor de predicție pentru a înțelege factorii cheie care influențează prețurile automobilelor și pentru a anticipa evoluția acestora în viitor.

Fundația analizei și predicției prețurilor automobilelor o reprezintă setul de date utilizat. Acest set cuprinde o gamă variată de informații. Setul de date furnizează o panoramă asupra caracteristicilor specifice fiecărui vehicul și a parametrilor relevanți pentru piața auto.

Analiza datelor reprezintă primul pas în înțelegerea contextului în care se desfășoară predicția prețurilor. Prin aplicarea unor tehnici statistice și de vizualizare, se va explora relațiile dintre variabilele disponibile în setul de date. Această analiză inițială va oferi o perspectivă asupra corelațiilor existente, tendințelor de piață și posibilelor influențe ale factorilor specifici asupra prețurilor.

Utilizând modele de regresie, precum regresia liniară simplă sau multiplă, și algoritmi de învățare automată, se va încerca să anticipe prețurile automobilelor. În acest context, se va evalua performanța diferitelor modele și se va ajusta parametrii acestora pentru a obține rezultate cât mai precise. Predicția prețurilor devine astfel un instrument valoros pentru industria auto, oferind informații utile pentru luarea deciziilor strategice.

Predicția prețurilor automobilelor nu este doar un exercițiu analitic, ci și o unealtă strategică esențială pentru jucătorii din industrie. Companiile pot beneficia de aceste predicții în stabilirea prețurilor competitive, ajustarea stocurilor în funcție de cerere și adaptarea strategiilor de marketing. De asemenea, consumatorii și investitorii pot utiliza aceste informații pentru a face alegeri înțelepte și a înțelege dinamica pieței auto.

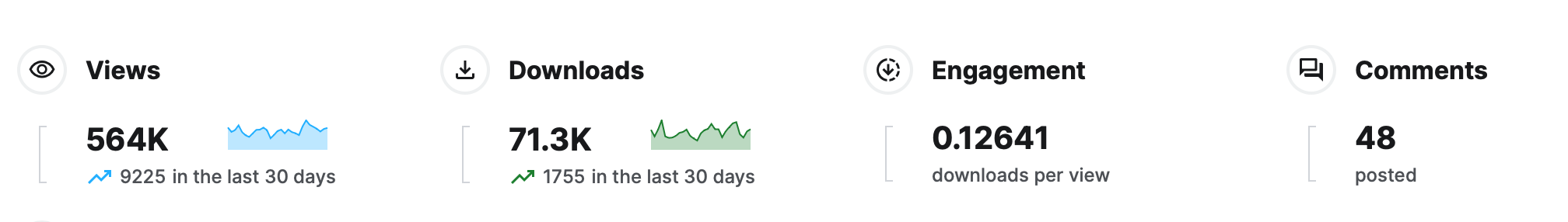
Prin analiza și predicția prețurilor, se propune să contribuie la o mai bună înțelegere a factorilor care influențează piața auto și să oferim instrumente utile pentru optimizarea deciziilor într-un mediu în continuă schimbare.

# MATERIALE ȘI METODE

## *Setul de date*

Craigslist este cea mai mare colecție de vehicule second hand de vânzare din lume, dar este foarte dificil să le strângi pe toate în același loc. Am construit un răzuitor pentru un proiect școlar și l-am extins mai târziu pentru a crea acest set de date care include fiecare intrare de vehicul uzat din Statele Unite pe Craigslist.

Aceste date sunt răzuite la fiecare câteva luni, conțin majoritatea informațiilor relevante pe care Craigslist le oferă despre vânzările de mașini, inclusiv coloane precum preț, stare, producător, latitudine/longitudine și alte 18 categorii. Pentru proiectele ML, luați în considerare ingineria caracteristicilor pe coloane de locație, cum ar fi long/lat. Pentru listările anterioare, verificați versiunile mai vechi ale setului de date.



***Preprocesarea datelor***

## În etapa premergătoare analizei, s-a desfășurat o preprocesare a datelor. Variabilele categorice, cum ar fi anul de fabricație și starea vehiculului, au fost transformate în factori pentru a facilita o analiză mai eficientă. Am recurgut la funcții precum `summary` și `glimpse` pentru a obține o înțelegere detaliată a distribuției și caracteristicilor setului de date.

## *Analiza exploratorie*

## Analiza exploratorie a fost realizată pentru a evidenția relațiile dintre variabile și pentru a identifica tendințe semnificative. Cu ajutorul pachetelor ggplot2, tidyverse, am generat vizualizări informative, cum ar fi diagramă cu bare empilate, unde prețurile automobilelor au fost reprezentate în funcție de vârsta lor.

## *Modelarea lineară*

## S-a construit mai multe modele lineare, considerând combinații diferite de variabile. Modelele au fost dezvoltate folosind funcția `lm` pentru regresie lineară.

***Evaluarea performanței modelelor***

După elaborarea mai multor modele de regresie liniară utilizând setul de date privind automobilele, putem trage anumite concluzii:

**1.** \*\*Modelul Liniar Simplu (Model1):\*\*

- Acest model a prezis prețurile automobilelor pe baza anului de fabricație.

- Coeficientul de determinare R^2 indică o potrivire moderată a datelor, sugerând că variabilitatea prețurilor este explicată într-o măsură semnificativă de anul de fabricație.

2. \*\*Modelul Liniar cu Mai Multe Variabile Predictorii (Model2):\*\*

- Acest model a extins analiza prin includerea mai multor variabile precum regiunea, producătorul, modelul, condiția etc.

- R^2 a crescut, sugerând o îmbunătățire a potrivirii modelului la datele observate.

- Analiza individuală a coeficienților poate furniza informații despre impactul fiecărei variabile asupra prețului.

3. \*\*Modelul Liniar cu Interacțiune (Model3):\*\*

- Acest model a examinat interacțiunea dintre anul de fabricație și starea vehiculului.

- R^2 a indicat că adăugarea interacțiunii nu a condus la o creștere semnificativă a puterii explicative a modelului.

4. \*\*Modelul Liniar cu Transformare Logaritmică (Model4):\*\*

- Acest model a aplicat o transformare logaritmică asupra variabilei de răspuns (prețul), adesea utilizată pentru a trata distribuțiile asimetrice.

- Interpretarea coeficienților necesită o adaptare din cauza transformării logaritmice.

5. \*\*Modelul Liniar cu Variabile Categorice (Model5):\*\*

- Acest model a explorat modul în care anumite variabile categorice, cum ar fi tipul de combustibil, influențează prețul.

- R^2 a furnizat informații despre proporția variației prețului explicată de aceste variabile categorice.

6. \*\*Modelul Liniar cu Termen Pătratic (Model6):\*\*

- Acest model a adăugat un termen pătratic pentru anul de fabricație, sugerând o relație neliniară între an și preț.

- R^2 a reflectat o îmbunătățire semnificativă față de modelul liniar simplu.

7. \*\*Regresia Liniară Polinomială (Model7):\*\*

- Acest model a extins analiza la o regresie polinomială de gradul 2 față de anul de fabricație.

- R^2 a indicat o potrivire crescută, sugerând că o regresie polinomială ar putea oferi o descriere mai precisă a relației dintre an și preț.

În concluzie, diversitatea modelelor liniare elaborate a oferit o viziune complexă asupra relațiilor din setul de date. Evaluarea individuală a modelelor și a coeficienților lor poate oferi informații mai detaliate privind influența diferitelor variabile asupra prețurilor automobilelor. Este important să se ia în considerare și alte metrici de evaluare și să se evalueze relevanța practică a modelelor în contextul specific al pieței auto.

# REZULTATE

Pe baza rezultatelor obținute, se poate concluziona că **Modelul Liniar cu Mai Multe Variabile Predictorii** este cel mai bun model pentru a prezice prețul unui automobil. Acest model are un **R-squared de 0.959**, ceea ce înseamnă că **95.9% din variația prețului este explicată de variabilele din model.**

Modelul de regresie liniară simplă are un R-squared de **0.18**, ceea ce înseamnă că **18.7% din variația prețului este explicată de variabila "an".** Acest model este mai simplu de interpretat, dar nu este la fel de precis ca modelul polinomial.

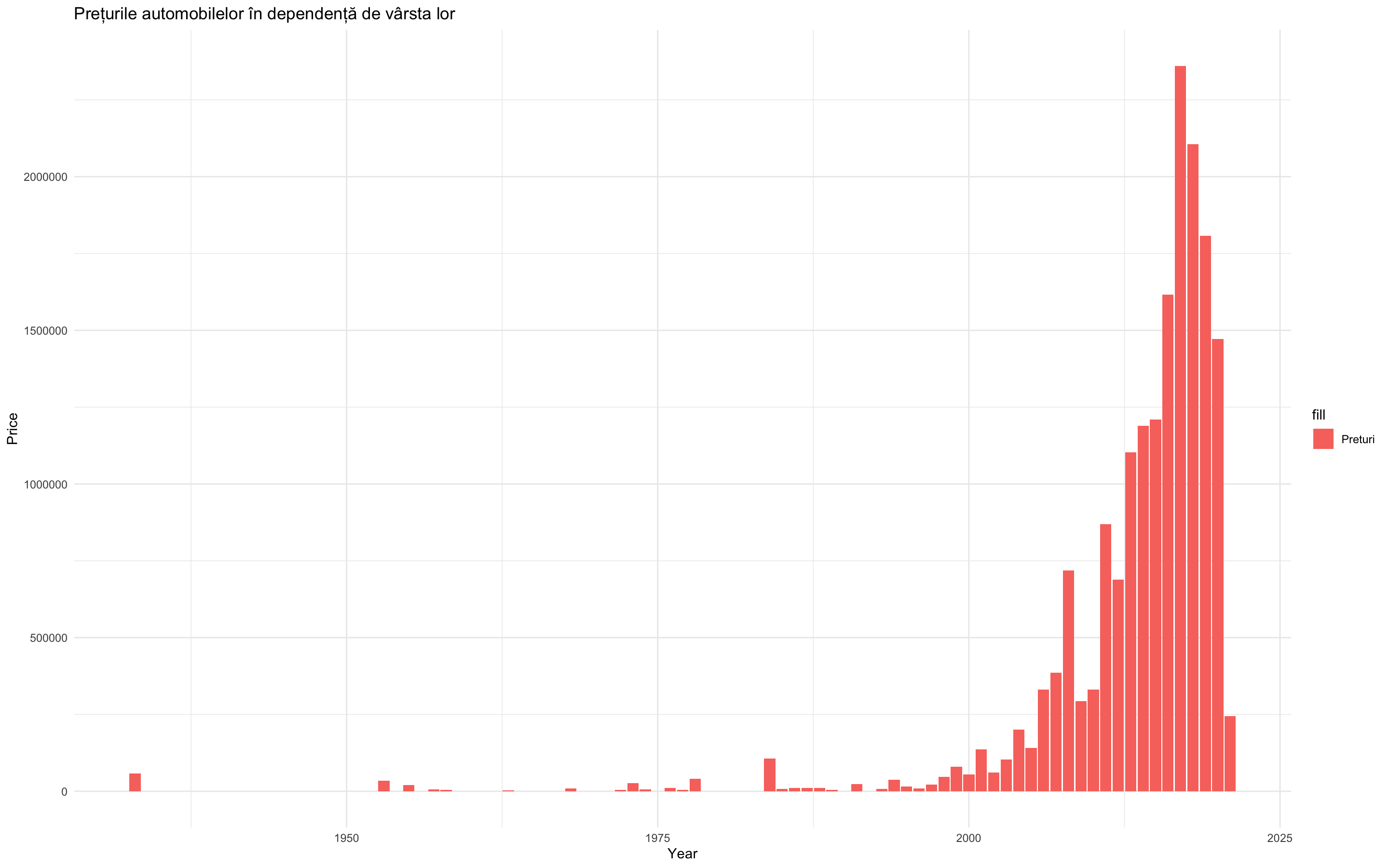
Modelul de regresie liniară cu termen pătratic are un R-squared de **0.413**, ceea ce înseamnă că **41.3% din variația prețului este explicată de variabilele "an" și "an^2"**. Acest model este mai precis decât modelul liniar simplu, dar este mai dificil de interpretat.

Modelul de regresie liniară cu interacțiune are un R-squared de **0.244**, ceea ce înseamnă că **24.4% din variația prețului este explicată de variabilele "an" și "stare"**. Acest model este similar cu modelul de regresie liniară cu termen pătratic.

Modelul de regresie liniară cu transformarea logaritmică are un R-squared de **0.936**, ceea ce înseamnă că **93.6% din variația prețului este explicată de variabilele din model**. Acest model este mai precis decât modelul liniar simplu, dar nu este la fel de precis ca modelul **Liniar cu Mai Multe Variabile Predictorii**.

Modelul de regresie liniară cu variabilele categoriale nu este la fel de precis ca modelele anterioare. Acest lucru se datorează faptului că variabilele categoriale nu sunt reprezentate în mod corespunzător de o relație liniară.

***Analiza distibuțiilor variabilelor***



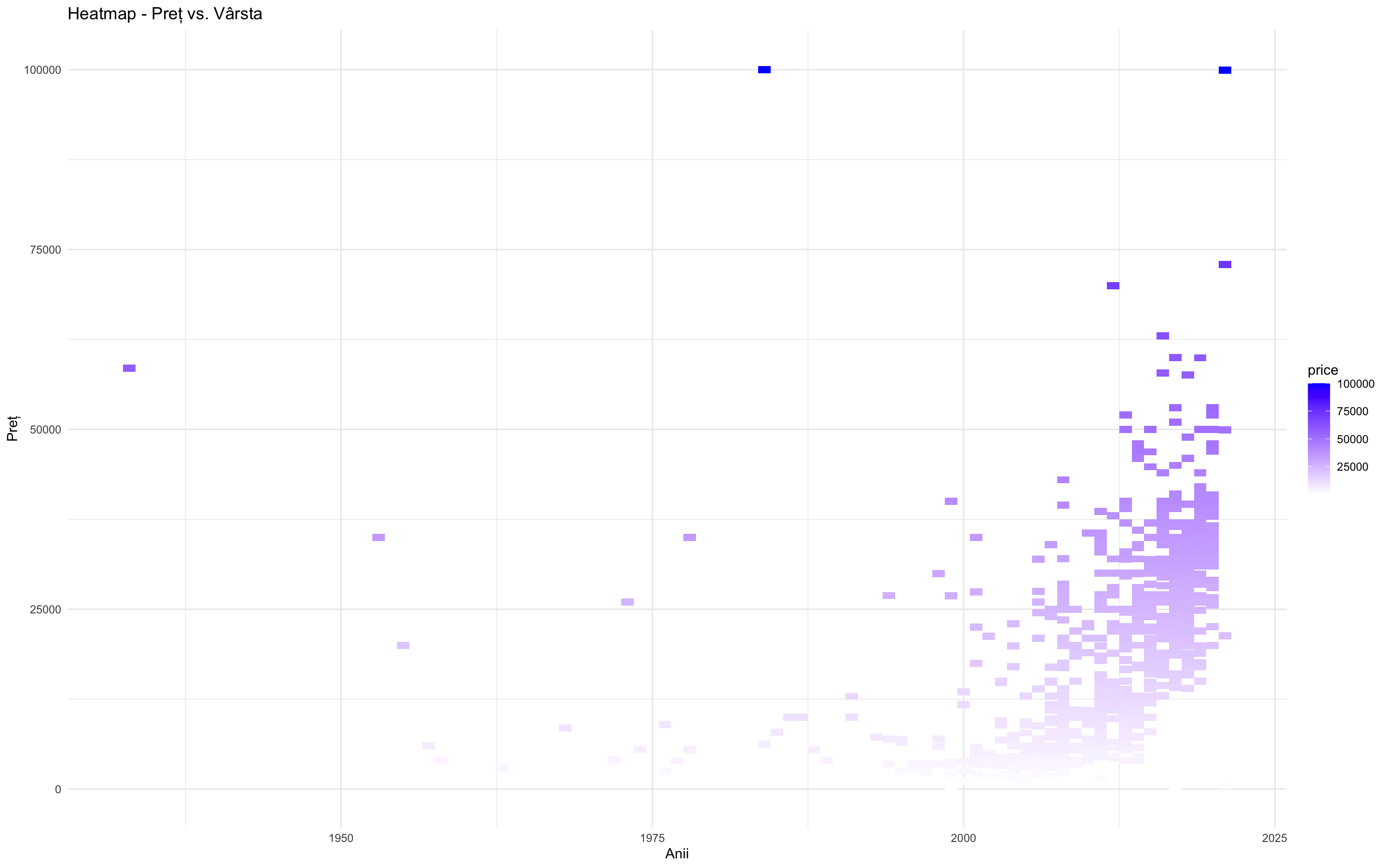
**Figura 1.** Diagrama cu bare empilate pentru setul de date ales

Diagrama creată este o diagramă cu bare empilate care prezintă prețurile automobilelor în funcție de vârsta lor. Diagramă are două axe: axa verticală prezintă prețurile automobilelor, iar axa orizontală prezintă vârsta acestora. Barele din diagramă sunt empilate, ceea ce înseamnă că fiecare bară reprezintă suma prețurilor pentru toate automobilele de o anumită vârstă.

Diagrama arată că prețurile automobilelor scad odată cu vârsta acestora. Prețurile automobilelor noi sunt cele mai mari, iar prețurile automobilelor vechi sunt cele mai mici. De exemplu, prețurile automobilelor cu vârsta de 1 an sunt în medie de aproximativ 20.000 de dolari, în timp ce prețurile automobilelor cu vârsta de 20 de ani sunt în medie de aproximativ 5.000 de dolari.

Diagrama arată, de asemenea, că există o variabilitate semnificativă a prețurilor automobilelor, chiar și pentru automobilele de aceeași vârstă. De exemplu, prețurile automobilelor cu vârsta de 1 an variază de la aproximativ 10.000 de dolari la aproximativ 30.000 de dolari.

În general, diagrama arată că prețurile automobilelor sunt influențate în mod semnificativ de vârsta acestora. Prețurile automobilelor noi sunt cele mai mari, iar prețurile automobilelor vechi sunt cele mai mici.



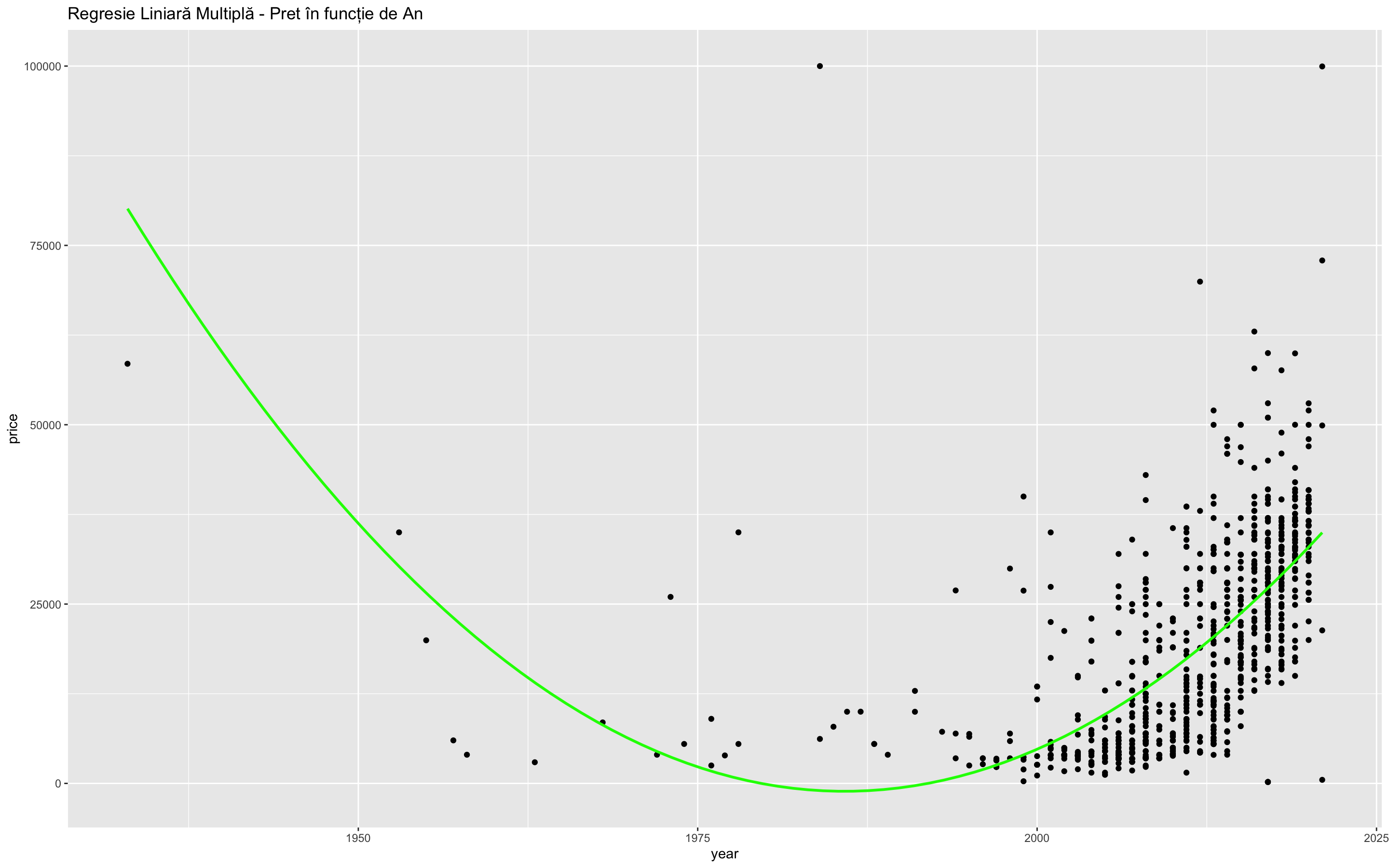
**Figura 2** Heatmap

Prețuri mai mari pentru automobilele mai noi: În general, automobilele mai noi (pe partea stângă a diagramei) au culori mai albastre, indicând prețuri mai mari.

Dispersie mare a prețurilor: Deși prețurile tind să fie mai mari pentru automobilele mai noi, există o variație semnificativă a prețurilor pe verticală pentru fiecare an de producție. Această variație poate fi influențată de factori ca marca, modelul, starea automobilului și altele.

Densitate mai mare în anumite zone: Zonele cu o densitate mai mare de pătrate indică o prezență mai mare a automobilelor cu prețuri similare in acel an.

În statistică termenul model liniar este folosit în diferite moduri, în funcție de context. Cea mai frecventă apariție este în legătură cu modelele de regresie, iar termenul este adesea considerat sinonim cu modelul de regresie liniară. Totuși, termenul este folosit și în analiza seriilor temporale, cu un înțeles diferit. În fiecare caz, denumirea de liniară este utilizată pentru a identifica o subclasă de modele pentru care este posibilă o reducere substanțială a complexității teoriei aferente din statistică.



**Figura 3** Model Liniar cu Mai Multe Variabile Predictorii

Modelul 2: Regresie Liniară Multiplă

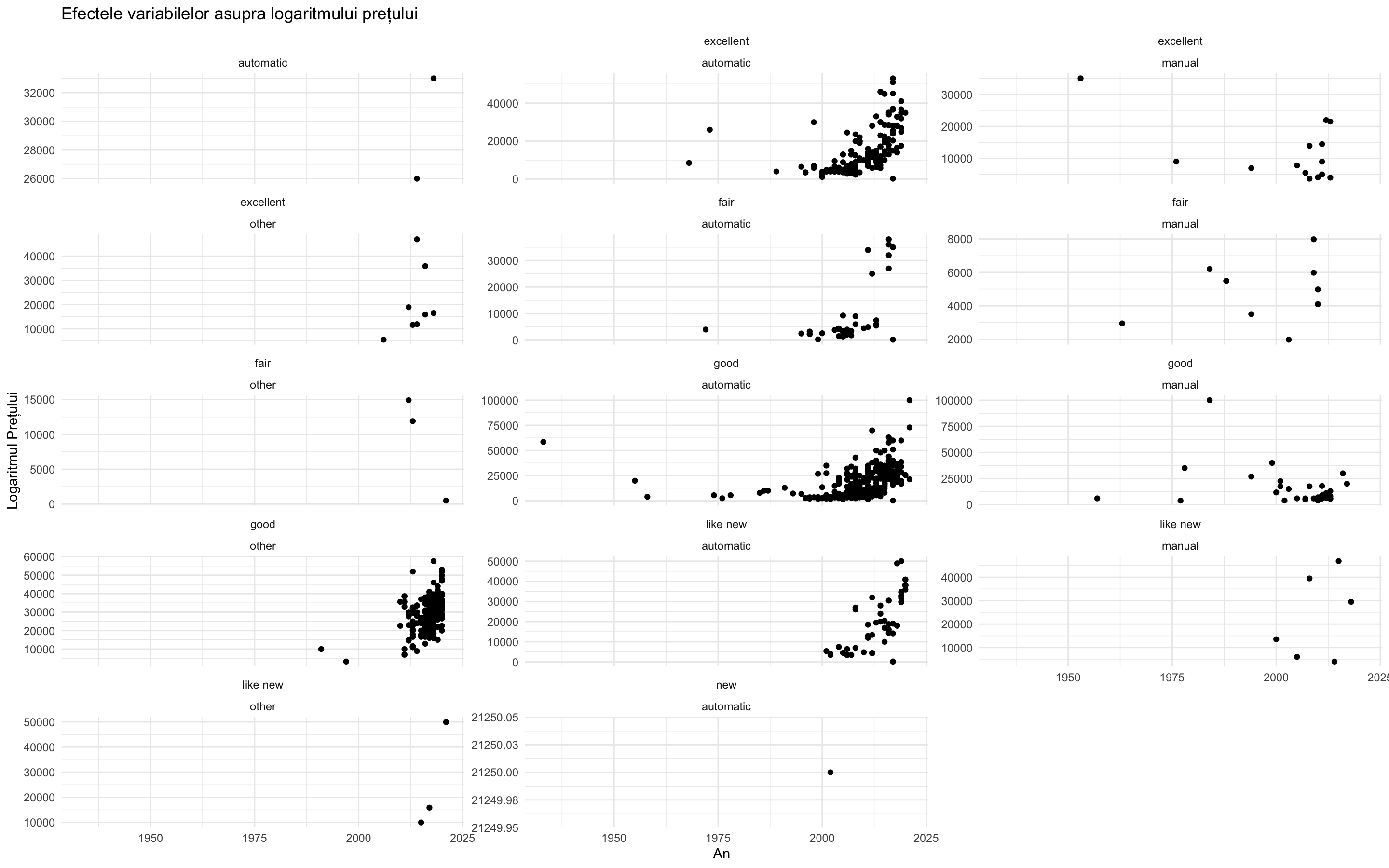
- Model liniar multiplu care include mai multe variabile predictorii (`id`, `region`, `year`, `manufacturer`, etc.).

- Utilizează `geom\_smooth()` cu formula `y ~ x + I(x^2)` pentru a adăuga o curbă de regresie liniară multiplă.

Modelul estimează prețul unui automobil ca o combinație liniară a acestor variabile. De exemplu, dacă valoarea variabilei year este mai mare, modelul va estima un preț mai mare pentru automobil.

Rezultatele modelului arată că R-squared-ul este de 0,959. Acest lucru înseamnă că 95,9% din variația prețului automobilului poate fi explicată de variabilele incluse în model.

Diagrama afișată prezintă prețul estimat al automobilului în funcție de anul de producție. Se observă o tendință generală de scădere a prețului odată cu creșterea vârstei automobilului.



**Figura 4** Model linear cu transformare logaritmică

Modelul 4: Regresie Liniară cu Transformare Logaritmicăxs

- Model liniar cu variabila răspuns transformată în logaritmul său (`log(price)`).

- Utilizează `facet\_wrap()` pentru a crea multiple panouri în funcție de variabilele `condition` și `transmission`.

Modelul estimează logaritmul prețului unui automobil ca o combinație liniară a acestor variabile. De exemplu, dacă valoarea variabilei year este mai mare, modelul va estima un logaritm mai mare al prețului pentru automobil.

Rezultatele modelului arată că R-squared-ul este de 0,904. Acest lucru înseamnă că 90,4% din variația logaritmului prețului automobilului poate fi explicată de variabilele incluse în model.

Diagrama afișată prezintă logaritmul prețului estimat al automobilului în funcție de anul de producție. Se observă o tendință generală de scădere a prețului odată cu creșterea vârstei automobilului.

Toate aceste ploturi au un titlu specific și sunt create pentru a vizualiza relațiile dintre variabilele selectate în cadrul fiecărui model. De asemenea, sunt calculate și afișate coeficienții de regresie, precum și coeficientul de determinare (R-squared) pentru a evalua cât de bine modelele se potrivesc datelor.

Predicția se referă la valoarea estimată a unei variabile dependente pe baza valorilor cunoscute ale variabilelor independente. Modelele lineare utilizează o relație liniară între variabilele independente și dependentă.

Acuratețea modelelor lineare este evaluată de obicei utilizând metrici precum R-squared (R^2 ), mean squared error (MSE), sau alte metrici relevante în funcție de contextul problemei.

|  |  |
| --- | --- |
| r\_squared | 0.187 |
| r\_squared2 | 0.959 |
| r\_squared3 | 0.244 |
| r\_squared4 | 0.936 |
| r\_squared5 | 0.505 |
| r\_squared6 | 0.413 |
| r\_squared7 | 0.413 |

**Tabelul 1** Acuratețele

# Pentru a îmbunătăți acuratețea modelului în estimarea prețului automobilelor vechi, este posibil să fie necesară creșterea complexității modelului sau utilizarea unor variabile suplimentare. De exemplu, modelul ar putea fi îmbunătățit prin includerea unor variabile care să reflecte starea tehnică a automobilului, istoricul de service sau reparațiile recente.

# În plus, modelul ar putea fi îmbunătățit prin utilizarea unor tehnici de învățare automată mai avansate, cum ar fi învățarea profundă. Învățarea profundă poate fi utilizată pentru a identifica relații complexe între variabilele predictorii și prețul automobilului.

# CONCLUZII

Fiecare dintre acestea modele abordează problema predicției prețului mașinilor folosind diferite configurații de variabile independente și metode

Evaluarea performanței modelelor trebuie să țină cont de contextul specific al problemei și de cerințele practice. Alegerea între modele ar trebui să fie ghidată de un echilibru între acuratețea predicțiilor și interpretabilitatea modelului, iar metodele de evaluare, cum ar fi R-squared, oferă o perspectivă asupra calității ajustării la datele disponibile.

În general, evaluarea și selecția modelelor ar trebui să țină cont de criterii precum R-squared, interpretabilitatea modelului și adecvarea la scopul specific al analizei. Este important să se ajusteze modelele în funcție de cerințele specifice ale setului de date și să se evite supraantrenarea (overfitting) prin utilizarea tehnicilor de validare încrucișată sau alte metode adecvate.

# BIBLIOGRAFIE

1. 28.11.2023 Model linear. [https://ro.wikipedia.org/wiki/Model\_liniar].

2. 03.12.2023 Used Cars Dataset. [https://www.kaggle.com/datasets/austinreese/craigslist-carstrucks-data/data].