**Construirea unui model predictiv pentru analiza vinzarii**

**Imobiliarelor**

Lazar Sergiu

Universitatea Tehnica a Moldovei

sergiu.lazar@iis.utm.md

Această cercetare a implicat utilizarea diferitelor tehnici statistice în limbajul de programare R și în mediul de dezvoltare RStudio pentru a analiza performanța vânzărilor de proprietăți imobiliare. Scopul studiului a fost să exploreze relațiile dintre variabilele implicate în procesul de vânzare a imobilelor, să facă predicții cu privire la evoluția pieței imobiliare, pentru o cunoastere mai buna si o intelegere profunda a comportamentului pietii de imobiliare.

Pentru analiza performanței vânzărilor imobiliare, s-au utilizat mai multe tehnici statistice, inclusiv regresia liniară pentru a evalua relațiile dintre diferitele variabile care pot influența prețurile și viteza de vânzare a proprietăților. S-au aplicat, de asemenea, metode de selectare, cum ar fi clasificarea Stepwise, si calculul corelatiei pentru a analiza importanta variabilelor si a le selecta pe cele mai relevante pentru procesul de strudiu. Evaluarea performanței modelelor a inclus analiza R-squared și a erorii standard reziduale, oferindu-ne o perspectivă asupra acurateței și fiabilității predicțiilor. Am comparat, de asemenea, predicțiile modelului cu valorile reale pentru a evalua capacitatea acestuia de a face predicții precise în practică..

Acest studiu a avut ca rezultat evidențierea relațiilor semnificative între variabilele implicate în piața imobiliară și a furnizat informații utile pentru înțelegerea și îmbunătățirea procesului de vânzare și cumpărare a proprietăților imobiliare. Prin această analiză detaliată, am căutat să optimizăm modelul pentru a reflecta cât mai fidel realitățile pieței imobiliare, având în vedere variația mare a prețurilor și complexitatea factorilor care le influențează

În final, este esențial să recunoaștem că, în ciuda avansurilor în modelarea statistică, predicția prețurilor imobiliare rămâne o provocare datorită naturii dinamice și multifactoriale a pieței. Prin urmare, modelele noastre ar trebui văzute ca instrumente care oferă perspective și ghiduri, dar care necesită interpretare prudentă și în contextul unei cunoașteri aprofundate a pieței locale și a trendurilor actuale.

**Introducere:**

Piața imobiliară este un domeniu dinamic și complex, unde numeroși factori influențează prețurile proprietăților. Acest articol urmărește să exploreze și să analizeze diverse caracteristici ale proprietăților imobiliare, cum ar fi numărul de dormitoare, băi, suprafața locuibilă, și altele, pentru a înțelege cum acești factori contribuie la stabilirea valorilor de piață.

Prin utilizarea unui set de date care cuprinde o varietate de proprietăți imobiliare, se propune a se efectua o analiză detaliată folosind limbajul de programare R. Aceasta va include atât analize descriptive, pentru a capta o imagine clară a stării actuale a pieței, cât și modele predictive, pentru a încercarea de a anticipa tendințele de preț bazate pe caracteristicile proprietăților.

Aceasta analiza se va axa pe tehnici statistice avansate, incluzând regresia liniară, clasificarea, analiza cu arbori de decizie, clusteringul K-means și analiza de corelație. Aceste metode permit explorarea relației complexe dintre variabile și identificarea factorii cheie care influențează prețurile imobiliare.

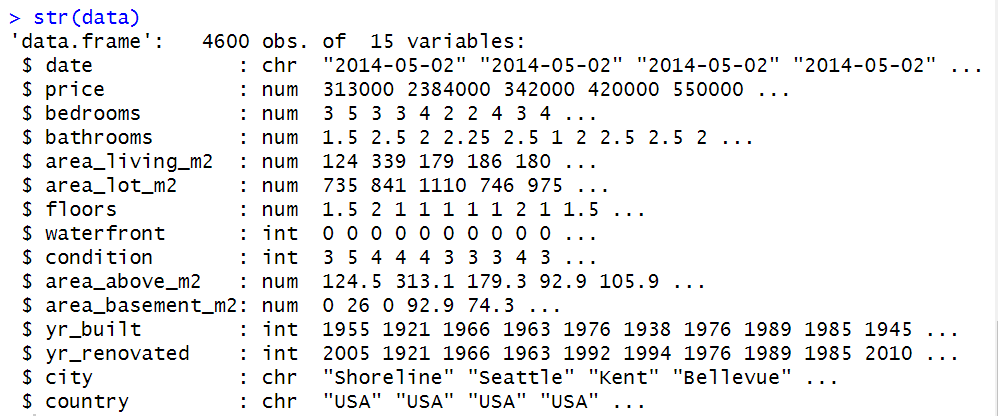
Scopul acestui articol este de a oferi o perspectivă cuprinzătoare asupra factorilor care modelează piața imobiliară, utilizând analiza de date pentru a aduce claritate într-un domeniu adesea perceput ca fiind neclar și imprevizibil.Acesta se concentreaza pe extragerea de înțelesuri valoroase din date, contribuind astfel la o mai bună înțelegere a dinamicii pieței imobiliare.

În următoarele secțiuni, se va explora setul de date, se va efectua o analiză descriptivă pentru a înțelege distribuția și caracteristicile principale ale datelor. In final in baza analizei vor fi construite modele pentru a investiga și prezice prețurile proprietăților.

**Explorarea Setului de Date:**

Am ineput prin a examina setul de date pentru a înțelege structura și caracteristicile principale ale acestuia. Acest pas va include o verificare a calității datelor (cum ar fi lipsa valorilor, incoerențe) și o descriere sumară a datelor :

* Verificarea Structurii Datelor: Examinearea structuri setului de date pentru a înțelege tipurile de variabile pe care le conține și formatul acestora.



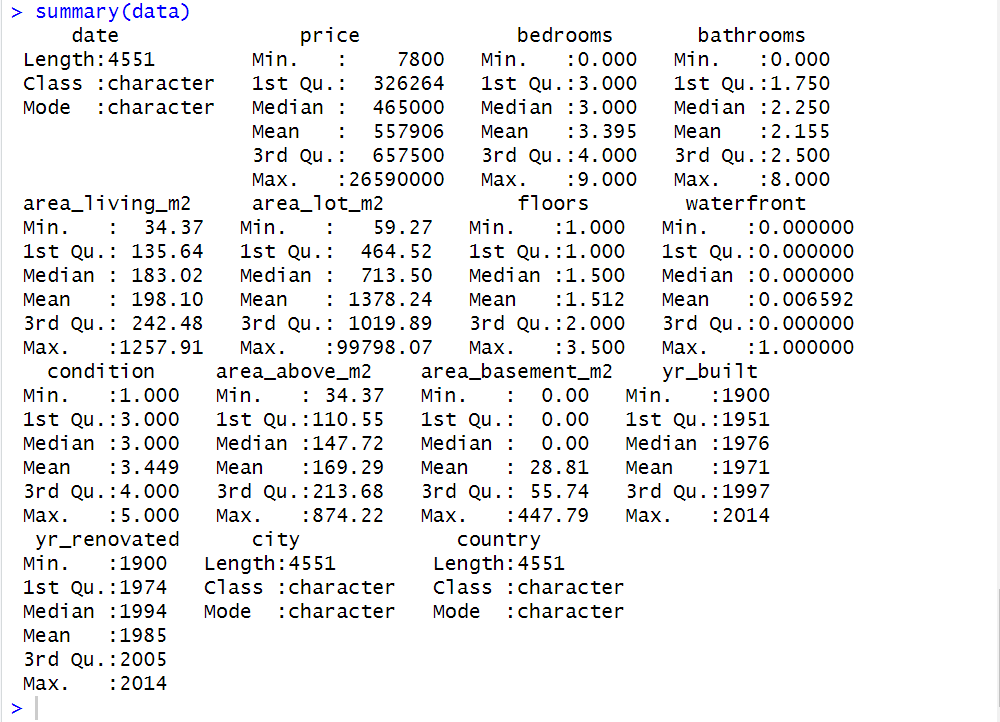
**Figura 1.** Structura generala a datelor

* Identificarea Valorilor Lipsă sau Anormale: Verificarea prezenței valorilor lipsă sau a datelor care par anormale (de exemplu, valori extreme sau incoerențe).



**Figura 2.** Verificarea valorilor lipsa

* Descrierea Sumară a Datelor: Caluclarea statistici sumare, cum ar fi media, mediana, deviația standard, pentru a obține o imagine de ansamblu asupra distribuțiilor variabilelor.



**Figura 3.** Statistica sumara a datelor

**Analiza Descriptivă:**

S-au realizat analize statistice pentru a descrie caracteristicile principale ale datelor. Acest lucru include calculul de statistici de bază, cum ar fi media, mediana, deviația standard, și altele. Aceste statistici vor oferi o imagine clară a distribuției fiecărei variabile din setul de date. De asemenea, au fost incluse și vizualizări (plot-uri) pentru a ajuta să înțelegem mai bine distribuțiile și relațiile dintre variabile.

Analiza statistica:

**Preț**: Interval de la 7,800 la 1,150,000, cu o medie de 482,786.

**Dormitoare**: Interval de la 0 la 9 dormitoare, cu o medie de 3.34.

**Băi**: Interval de la 0 la 5.75 băi, cu o medie de 2.077.

**Suprafața Locuibilă** (m²): Interval de la 34.37 m² la 398.55 m², cu o medie de 185.51 m².

**Suprafața Terenului** (m²): Interval de la 59.27 m² la 99,798.07 m², cu o medie de 1,312.26 m².

**Etaje**: Interval de la 1 la 3.5 etaje, cu o medie de 1.49 etaje.

**Waterfront** (vedere la apă): Interval de la 0 (nu) la 1 (da), cu o medie de 0.003521.

**Condiție**: Interval de la 1 la 5, cu o medie de 3.446.

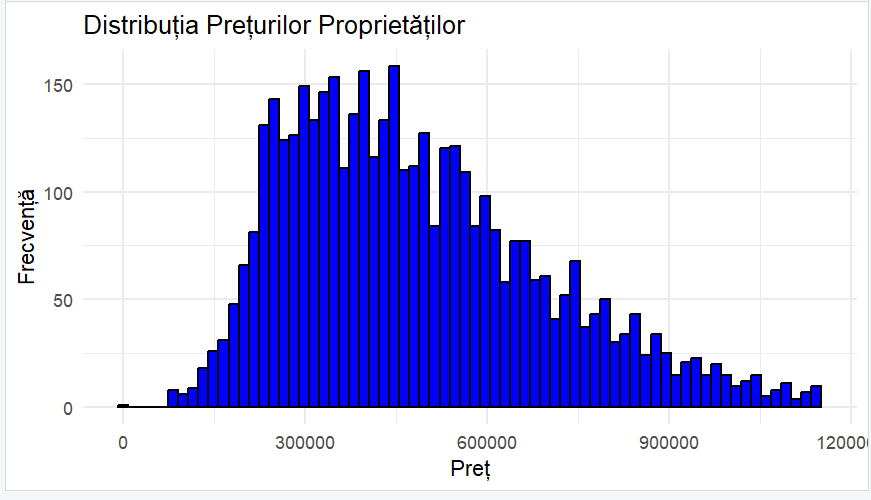
**Suprafața Deasupra** (m²): Interval de la 34.37 m² la 398.55 m², cu o medie de 159.81 m².

**Suprafața Subsol** (m²): Interval de la 0 m² la 195.1 m², cu o medie de 25.7 m².

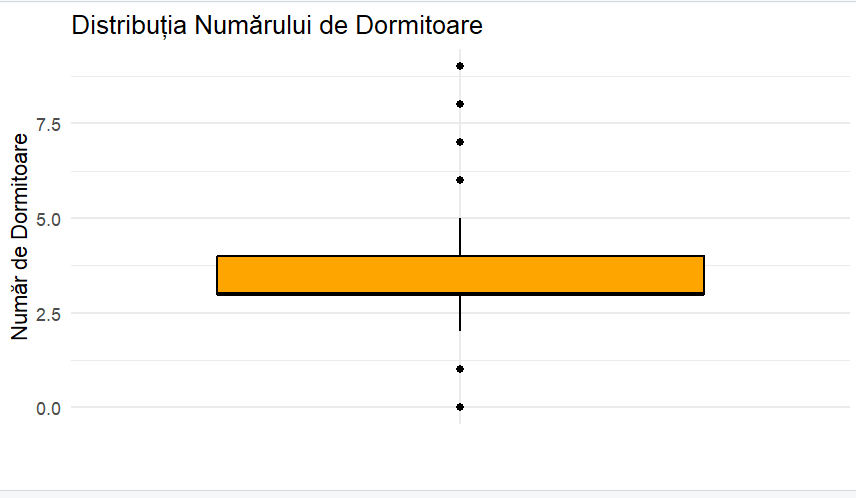
**Anul Construirii**: Interval de la 1900 la 2014, cu o medie de 1971.

**Anul Renovării:** Interval de la 1900 la 2014, cu o medie de 1985.

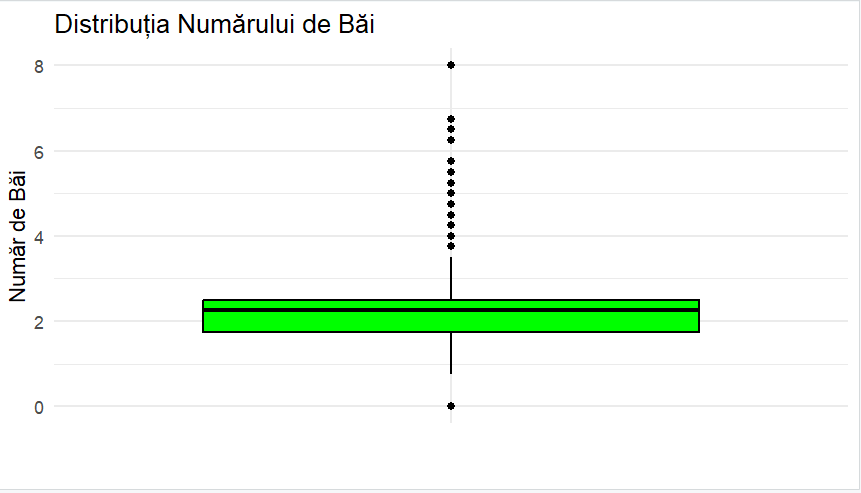
Vizualizarea unor variabile:



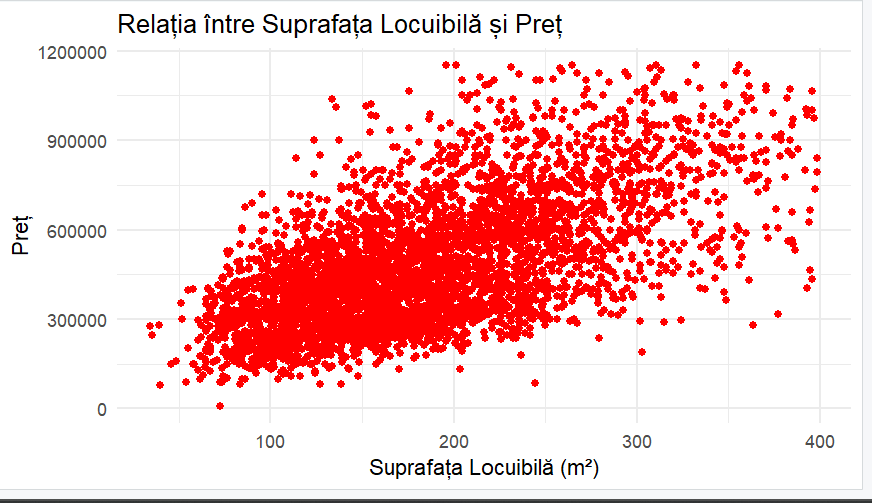
**Figura 4.** Distributiapreturilor



**Figura 5.** Distributia nr. dedormitoare



**Figura 6.** Distributia nr. debai



**Figura 7.** Relatia dintre suprafata si pret

**Analiza Corelațiilor și Relațiilor:**

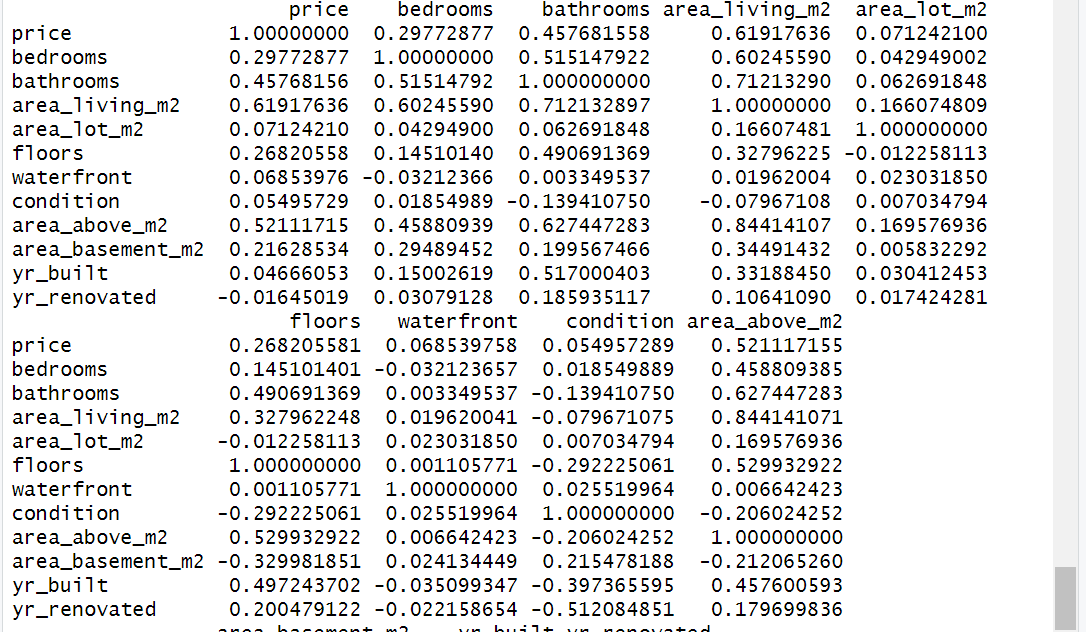
În această etapă a analizei, a fost pus accentul pe descoperirea și interpretarea relațiilor dintre diferite variabile ale setului de date. Scopul este de a înțelege cum interacționează diferite caracteristici ale proprietăților imobiliare, precum suprafața locuibilă, numărul de dormitoare, băi și alte aspecte, cu prețul acestora. Prin explorarea acestor relații, se pot identifica factorii cheie care influențează valoarea de piață a proprietăților și cum acești factori interacționează între ei.

Primul pas a fost calcularea coeficienților de corelație Pearson, care ofera o măsură statistică a gradului de corelație liniară între perechi de variabile. Această abordare ajuta la identificarea relațiilor semnificative .

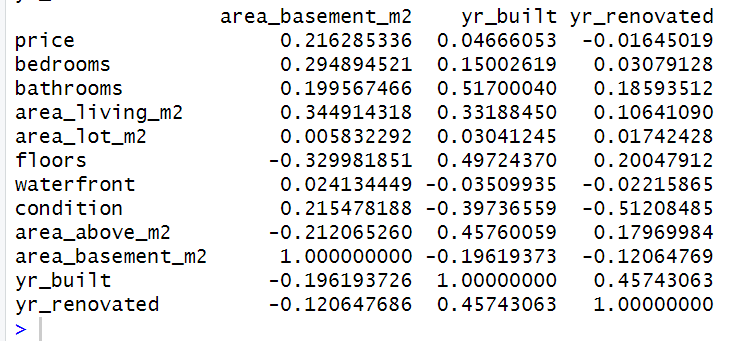
În plus, s-au folot vizualizări de date, cum ar fi heat maps, pentru a ilustra aceste corelații. Aceste grafice permit observarea vizuala a modelelor și tendințelor din date, oferind o perspectivă intuitivă asupra modului în care variabilele sunt legate unele de altele.

Această analiză detaliată, faciliteaza procesul de a trage concluzii valoroase care vor usura înțelegerea pieței de imobiliare, permițându-ne să facem predicții mai informate și să identificăm oportunități potențiale de investiții.

Matricea de corelatie:



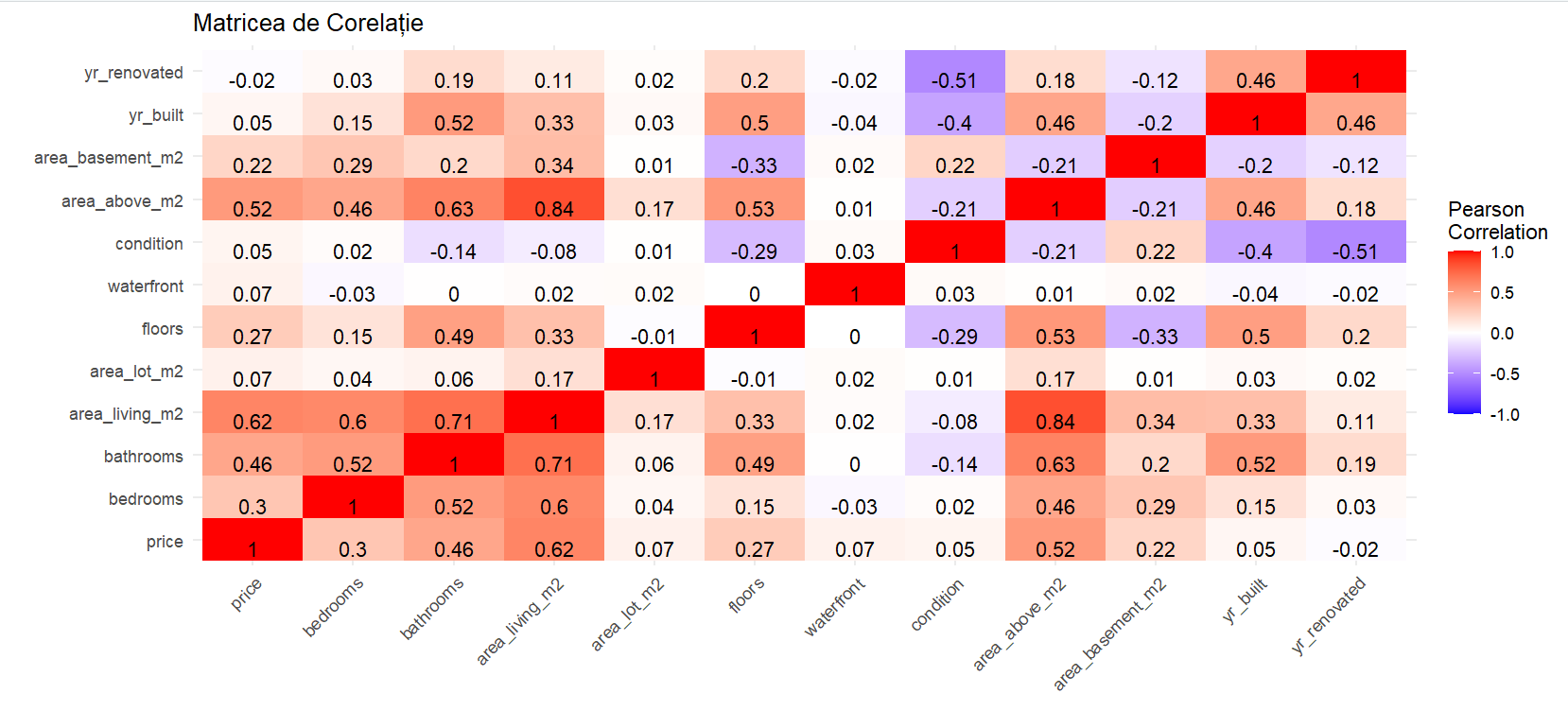
**Figura 8.** Matricea de corelatie partea I



**Figura 9.** Matricea de corelatie partea II

Pe baza matricii de corelatie observam:

* **Preț și Suprafața Locuibilă**: Există o corelație moderat puternică (aproximativ 0.62) între preț și suprafața locuibilă a proprietăților, sugerând că pe măsură ce suprafața locuibilă crește, crește și prețul, ceea ce este de așteptat în piața imobiliară.
* **Preț și Numărul de Dormitoare/Băi**: Există corelații pozitive, dar mai slabe, cu prețul: 0.30 pentru dormitoare și 0.46 pentru băi. Aceasta indică faptul că numărul mai mare de dormitoare și băi poate contribui la creșterea prețului, dar nu în aceeași măsură ca suprafața locuibilă.
* **Preț și Etaje**: Există o corelație pozitivă modestă (aproximativ 0.27) între preț și numărul de etaje, sugerând că proprietățile cu mai multe niveluri pot avea o valoare ușor mai mare, deși acest factor nu este la fel de influent ca suprafața locuibilă.
* **Alte Corelații**: Corelațiile mai slabe ale prețului cu factori precum vederea la apă (0.07), condiția proprietății (0.05) și anul construirii (0.05) sugerează că acești factori pot avea un impact mai puțin direct sau mai subtil asupra valorii de piață.



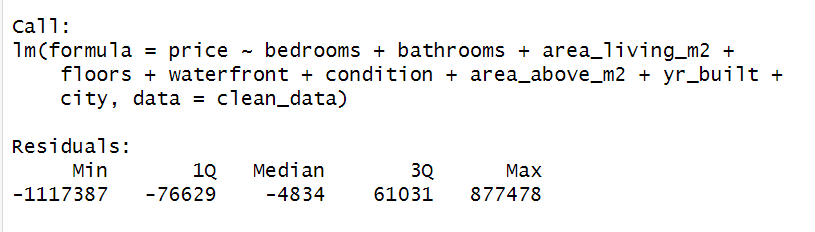
**Figura 10.** Heat map

**Modelare Predictivă:**

La aceasta etapa ne vom concentra asupra modelării predictive, un proces esențial pentru înțelegerea și anticiparea dinamicii prețurilor în piața imobiliară. Scopul principal al modelării predictive este de a construi un model statistico-matematic care să permită să prezicem prețurile proprietăților imobiliare bazându-ne pe o serie de caracteristici relevante.

1. *Selectarea Variabilelor*:

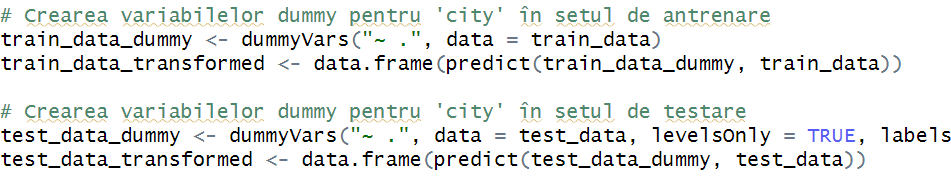
Inițial, a fost pus accent pe identificarea celor mai semnificative variabile care influențează prețurile. Prin aplicarea metodei stepwise regression, am reușit să selectăm un subset de variabile care au arătat cel mai mare potențial predictiv. Această etapă a fost crucială pentru a asigura că modelul nostru este atât eficient, cât și relevant din punct de vedere al interpretării.



**Figura 11.** Aplicarea stepwise

2. *Prelucrarea Datelor:*

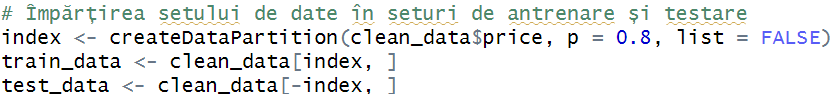
Am continuat cu prelucrarea setului de date, inclusiv codificarea variabilelor categorice, cum ar fi orașul, în variabile dummy. Acest pas a fost vital pentru a gestiona variabilele categorice cu multiple niveluri și pentru a transforma datele într-un format adecvat pentru modelare.



**Figura 12.** Prelucrarea datelor categoriale

3. *Împărțirea Setului de Date:*

În continuare, am împărțit setul de date în subgrupuri de antrenare și testare. Aceasta ne-a permis să antrenăm modelul pe un subset de date și să evaluăm performanța acestuia pe un set de date independent, asigurându-ne astfel că modelul poate generaliza dincolo de datele pe care a fost antrenat.



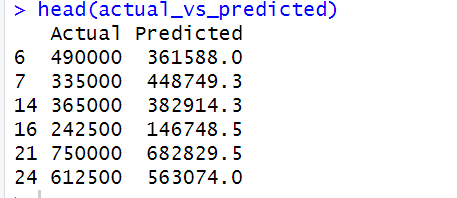
**Figura 13.** Divizarea Datelor

*4. Construirea Modelului de Regresie Liniară:*

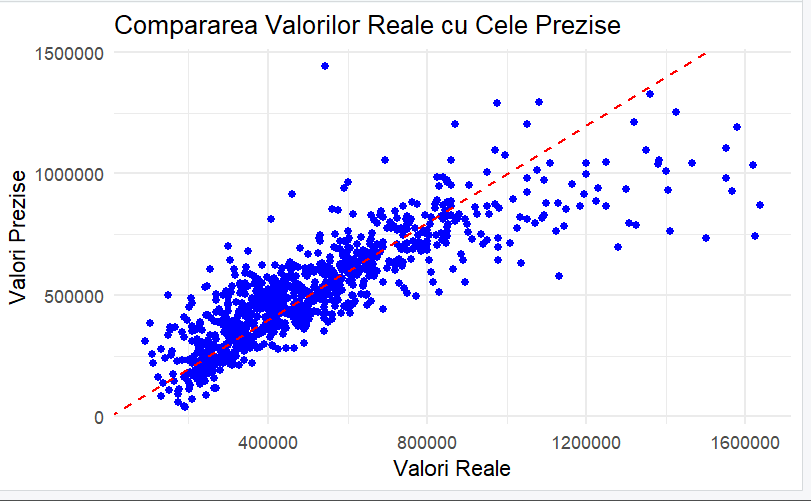
Utilizând variabilele selectate, am construit un model de regresie liniară. Acest model ne-a permis să analizăm relația dintre variabilele independente și prețul proprietăților, oferindu-ne o bază solidă pentru a face predicții.

*5. Evaluarea Modelului:*

În final, am evaluat performanța modelului prin compararea predicțiilor cu valorile reale din setul de testare. Acest pas a fost crucial pentru a evalua acuratețea și fiabilitatea modelului nostru predictiv.



**Figura 14.** Compararea rezultatelor reale cu cele prezise de model



**Figura 15.** Compararea grafica a rezultatelor reale cu cele prezise de model

Modelul de regresie liniară prezintă o performanță decentă, cu un R-squared de 0.6864, indicând că aproximativ 68.64% din variația prețului este explicată de variabilele modelului.Deși modelul are o capacitate bună de a explica variația prețurilor, comparația dintre valorile prezise și cele reale arată diferențe semnificative în unele cazuri. Acest lucru sugerează că, deși modelul este util, există loc pentru îmbunătățiri în acuratețea predicțiilor.

**Interpretarea Rezultatelor Regresiei**

1. ***Coeficienții Modelului:***

Bedrooms, Bathrooms și Area Living: Coeficienții pozitivi pentru variabile precum numărul de dormitoare (bedrooms), numărul de băi (bathrooms) și suprafața locuibilă (area\_living\_m2) indică o relație directă cu prețul. Astfel, pe măsură ce numărul de dormitoare sau băi crește sau suprafața locuibilă se mărește, prețul estimat al proprietății tinde să crească.

1. ***R-squared și Adjusted R-squared:***

Un R-squared de 0.6864 sugerează că aproximativ 68.64% din varianța prețurilor este explicată de variabilele incluse în model. Acest lucru indică o putere predictivă bună, dar nu perfectă, a modelului.

Adjusted R-squared de 0.6763 ajustează R-squared pentru numărul de predictorii din model, oferind o măsură mai realistă a calității fitului.

1. **Eroarea Standard Reziduală:**

O eroare standard reziduală de 148,500 indică mărimea tipică a deviației dintre valorile prezise de model și valorile reale ale prețului. Aceasta este o măsură a acurateței modelului, iar în acest caz, sugerează că modelul are o anumită precizie, dar și un grad de incertitudine în predicții.

1. **P-*value***:

Un p-value foarte mic (< 2.2e-16) pentru F-statistica indică faptul că modelul este statistic semnificativ și că există o relație între variabilele independente și preț.

1. **Interpretarea Practică:**

Comparând valorile prezise cu cele reale, observăm că există diferențe în anumite predicții. Acest lucru subliniază importanța unei analize atente și a interpretării rezultatelor în contextul real al pieței imobiliare.

1. ***Limitări și Considerații:***

Trebuie să luăm în considerare factori externi și limitările datelor, precum potențialele variabile neobservate sau erorile de măsurare, care pot influența acuratețea modelului.

Există posibilitatea ca modelul să fie supus overfitting-ului sau să nu captureze toate nuanțele și complexitățile pieței imobiliare.

**Concluzii:**

În concluzie, analiza aprofundată a datelor pieței imobiliare ne-a oferit posibilitatea de a dezvălui relații semnificative și de a construi modele care să prezică și să înțeleagă mai bine variația prețurilor proprietăților. Prin includerea unei game variate de variabile, ajustarea modelelor de regresie liniară, analiza reziduurilor și aplicarea diferitelor tehnici statistice, am obținut o perspectivă detaliată asupra factorilor care influențează valorile imobiliare.

Am constatat că variabile precum suprafața locuibilă, numărul de dormitoare și băi sunt semnificativ corelate cu prețul proprietăților. Analiza factorilor de inflație a varianței (VIF) a fost esențială în identificarea și gestionarea coliniarității în modelele noastre. De asemenea, am explorat relații complexe și am evaluat precizia modelelor noastre prin comparația directă a valorilor prezise cu cele reale**.**

Modelul oferă o perspectivă valoroasă asupra factorilor care influențează prețurile imobiliare și poate fi un instrument util în predicția prețurilor. Cu toate acestea, trebuie abordat cu o înțelegere clară a limitărilor sale și a naturii probabilistice a predicțiilor în analiza de date

**Disponibilitatea codului si datelor**

https://github.com/lazar006/Analiza-Datelor

**Biblliografie**

[1] RStudio Team (2023). RStudio IDE: A graphical user interface for R. RStudio

[2]<https://ourcodingclub.github.io/tutorials/intro-to-r/>

[3]https://www.intellspot.com/types-data-analysis/

[4]https://www.kaggle.com/datasets/shree1992/housedata/data