# Analiza dependentei dintre caracteristicile imobiliarelor si prețul lor

**Scurtu Corneliu, IA-211, III**

*Universitatea Tehnică a Moldovei, Facultatea Calculatoare, Inginerie și Micro-Electronică, Chișinău, Moldova, Informatica aplicată.*

**Abstra­ct**

În era digitală în continuă expansiune, analiza setului de date California Estate devine esențială pentru înțelegerea dinamicii pieței imobiliare și a comportamentului consumatorilor. Acest studiu se axează pe interpretarea detaliată a informațiilor din setul de date, aducând în prim-plan nu doar aspecte cantitative, ci și implicațiile sociale și economice ale schimbărilor din domeniul imobiliar.

Cu ajutorul tehnicilor avansate de analiză de date, cercetarea explorează nu numai preferințele cumpărătorilor, ci și caracteristicile distinctive ale proprietăților care atrag atenția. Se analizează modul în care aspecte precum dimensiunile locuinței, facilitățile disponibile și amplasarea geografică influențează deciziile de cumpărare. De exemplu, se investighează cum trendurile actuale în materie de design și utilități afectează cererea pentru anumite tipuri de proprietăți, conturând astfel profilul consumatorului modern.

Pe lângă aspectele tradiționale legate de prețuri și cerere, acest studiu aduce în discuție și importanța evoluțiilor tehnologice în imobiliare, explorând cum inovațiile în domeniul smart home și eficiența energetică influențează deciziile cumpărătorilor. Astfel, analiza se extinde pentru a oferi o perspectivă comprehensivă asupra pieței imobiliare, ilustrând cum caracteristicile imobilelor devin la fel de cruciale precum factorii socio-economici în modelarea preferințelor și comportamentului consumatorilor.

# Introducere

Setul de date California Estate reprezintă o comoară informațională cu privire la dinamica complexă a pieței imobiliare din California, reflectând schimbările și tendințele semnificative în ceea ce privește achizițiile de proprietăți.

Setul de date cuprinde informații esențiale despre diverse aspecte, inclusiv tipuri de proprietăți, caracteristici distinctive ale imobilelor, prețuri, localizare geografică și tendințe temporale. De la case rezidențiale la unități comerciale și terenuri, acest set de date oferă o privire exhaustivă asupra diversității pieței imobiliare din California.

Analiza propusă își propune să cartografieze schimbările în preferințele cumpărătorilor și să evalueze impactul diferiților factori asupra prețurilor și cererii. Prin utilizarea unor tehnici de analiză de date vom explora caracteristicile specifice ale imobilelor care au un impact semnificativ asupra deciziilor de achiziție.

Această analiză nu numai că va dezvălui tendințe și modele ascunse în datele imobiliare, dar va oferi și o înțelegere mai profundă a dinamicii pieței, contribuind astfel la dezvoltarea unor strategii mai informate pentru actorii implicați în domeniul imobiliar din California. Prin conectarea aspectelor cantitative cu poveștile subiacente din spatele cifrelor, vom contura un portret comprehensiv al pieței imobiliare, aducând lumină în complexitatea și evoluția acestui sector vital al economiei californiene.

# Materiale & Metode

**Set de Date**

Pentru realizarea acestei analize imobiliare am utilizat un set de date, disponibil pe platforma open-source Kaggle.com[1], care oferă informații actualizate despre piața imobiliară din California pentru primele 6 luni ale anului 2021. Acesta aduna toate listările de proprietăți imobiliare din această regiune. Cu un număr impresionant de 39 de coloane și 35,389 de rânduri, acest set de date oferă o perspectivă detaliată asupra diversității pieței imobiliare californiene.

Acest set de date reprezintă o resursă valoroasă pentru toți cei interesați de piața imobiliară din California, de la agenți imobiliari și dezvoltatori la cercetători și consumatori. Analiza acestor date oferă o înțelegere aprofundată a evoluției pieței imobiliare în primele 6 luni ale anului 2021, permițând observarea tendințelor de preț, preferințelor consumatorilor și altor factori critici care pot influența deciziile viitoare în domeniul imobiliar. Cu variabile numerice, categoriale și temporale, acest set de date permite o analiză comprehensivă și detaliată, contribuind la conturarea unui peisaj clar al pieței imobiliare din California.

Variabilele incluse în setul de date sunt categorisite în funcție de natura lor, acoperind aspecte esențiale precum prețul proprietății, anul construcției, date critice legate de vânzări, informații despre utilizatori, și detalii unice despre fiecare proprietate, cum ar fi localizarea și data postării anunțului.

**Preprocesarea Datelor**

Pentru a pregăti datele din setul meu de date în vederea analizei, am folosit limbajul de programare R[2] și diverse biblioteci, inclusiv openintro[3] și tidyverse[4]. După încărcarea setului de date, am efectuat o analiză sumară, concentrându-ne asupra tipurilor de date existente. Astfel am identificat, un număr semnificativ de rânduri cu valori eronate, fie egale cu zero sau extrem de mari, care nu se potriveau contextului. Am aplicat un filtru pentru a elimina aceste valori incoerente.

În procesul de curățare, am filtrat valorile egale cu zero în ceea ce privește livingArea și price, eliminând astfel incoerențele din setul de date. De asemenea, am exclus tipurile de proprietăți APARTMENT și LOT din coloana type, deoarece acestea nu conțineau suficiente date valide pentru analiză. Proprietățile de tip LOT, fiind terenuri fără construcții asociate, prezentau diferențe semnificative în ceea ce privește prețul față de celelalte tipuri de proprietăți, afectând astfel rezultatele analizei. Aceste etape de curățare și filtrare au fost esențiale pentru a asigura coerența și relevanța datelor în perspectiva analizei ulterioare.

**Analiza Exploratorie a Datelor**

Explorând diverse perspective, am investigat legăturile dintre diferitele caracteristici ale proprietăților imobiliare din California și modul în care acestea influențează performanța pe piață. Prin intermediul vizualizărilor, am examinat distribuția prețurilor de vânzare în funcție de atribute precum numărul de camere sau amplasarea geografică a imobilelor. Utilizând instrumente precum ggplot[4], am creat grafice interactive pentru a ilustra în mod clar aceste relații complexe. Pentru a oferi o sinteză a datelor, am adoptat abordări statistice, precum funcția summary, scoțând în evidență aspectele semnificative ale setului de date California Estate și furnizând informații utile pentru înțelegerea pieței imobiliare din această regiune.

**Modelarea și Evaluarea Performanței**

În procesul de modelare și evaluare a performanței, m-am confruntat cu provocări intrigante și decizii importante. Alegerea algoritmului potrivit pentru setul meu de date a fost o etapă crucială, implicând o analiză atentă a caracteristicilor și particularităților acestuia. Preprocesarea datelor a fost o altă etapă esențială, unde am aplicat diverse transformări pentru a asigura calitatea și relevanța informațiilor. Pe măsură ce am implementat și evaluat modelul, am folosit măsuri specifice precum R2 pentru a obține o înțelegere cuprinzătoare a performanței sale. Un aspect asupra căruia am pus o atenție deosebită a fost selecția de caracteristici, înțelegând că aceasta poate influența semnificativ rezultatul final al modelului.

Un aspect asupra căruia am pus o atenție deosebită a fost selecția de caracteristici, înțelegând că aceasta poate influența semnificativ rezultatul final al modelului. Pe parcursul acestui proces provocator, am înțeles importanța optimizării modelului, contribuind la atingerea unui nivel optim de generalizare. Alegerea modelului de regresie liniară a adus claritate și consistență în procesul de modelare și evaluare a performanței.

**Codul disponibil**

Tot codul pentru crearea modelului pentru regresie liniară și pentru prezicere este disponibil pe [GitHub](https://raw.githubusercontent.com/scorneliu/ad/main/regression.R)

# Rezultate

In urma procesării datelor a urmat procesul de analiza a acestora, am început prin crearea unui ggplot pentru a verifica cererea pentru anumite tipuri de apartamente dintre care am determinat faptul ca in California cererea cat si ofertele pentru apartamente sunt mici, majoritatea fiind interesați de prosperități pe pământ. Se observa in special o oferta ridicata pentru tipurile de proprietăți de tip SINGLE\_FAMILY (Figura 1). Evenimentele “Price change” și “Sold” sunt relativ rare pentru toate tipurile de locuințe, ceea ce sugerează că piața imobiliară din California este stabilă sau lentă. Există foarte puține apartamente și condominii listate pentru închiriere comparativ cu celelalte tipuri de locuințe.

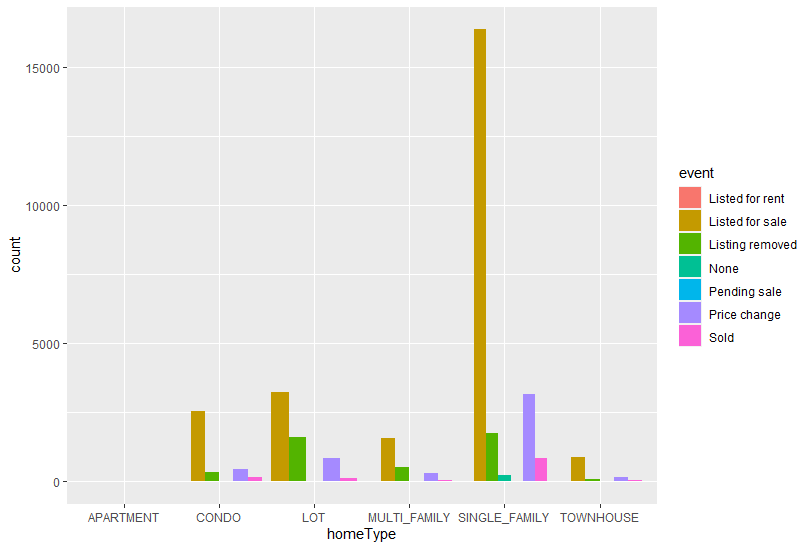


Figura 1 - Events count per tip de apartament

In acesta histograma am explorat tendințele preturilor in funcție de tipul de proprietate se poate observa media prețului pe tipul de proprietate cat si numărul de oferte pentru un anumit diapazon de preț. Se poate observa ca LOT și SINGLE\_FAMILY au cele mai multe proprietăți la prețuri mici, în timp ce MULTI\_FAMILY și TOWNHOUSE au o distribuție mai uniformă a prețurilor. MULTI-FAMILY are cele mai puține locuințe și un vârf la prețul de aproximativ 1,000,000. Graficul sugerează că există o cerere mare pentru locuințele ieftine, cum ar fi LOT și SINGLE\_FAMILY, și o ofertă limitată pentru locuințele scumpe, cum ar fi LOT și TOWNHOUSE. MULTI-FAMILY pare a fi un tip de locuință mai rar și mai costisitor.

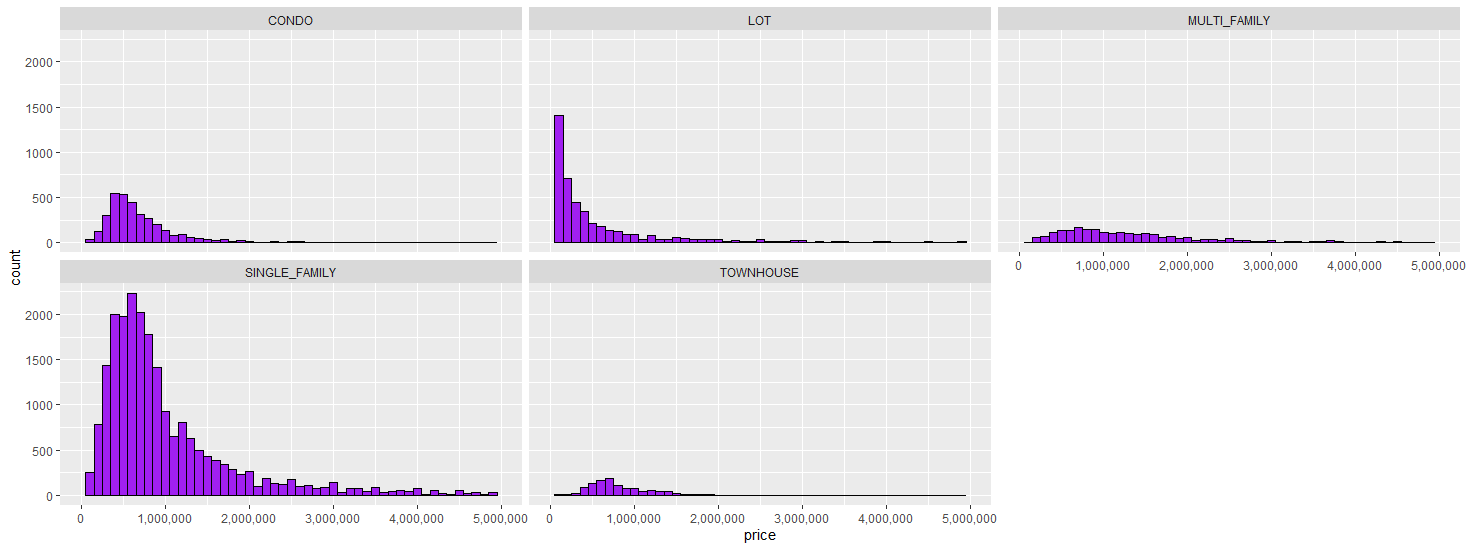


Figura 2 - Preturile după tipul de apartament

In urma analizezi histogramei cu preturile am determinat faptul ca proprietățile de tip LOT diferă foarte mult fata de celelalte tipuri de proprietăți deoarece acestea nu includ o construcție astfel am decis sa exclud aceste date pentru o mai buna acuratețe a modelului cat si o înțelegere mai buna a datelor. SINGLE\_FAMILY are cele mai multe vânzări, cu aproape 750 de unități vândute. TOWNHOUSE are cele mai puține vânzări, cu doar aproximativ 50 de unități vândute. CONDO are un număr moderat de vânzări, cu aproximativ 250 de unități vândute, în timp ce LOT are mai puține vânzări decât CONDO, cu puțin peste 100 de unități vândute.

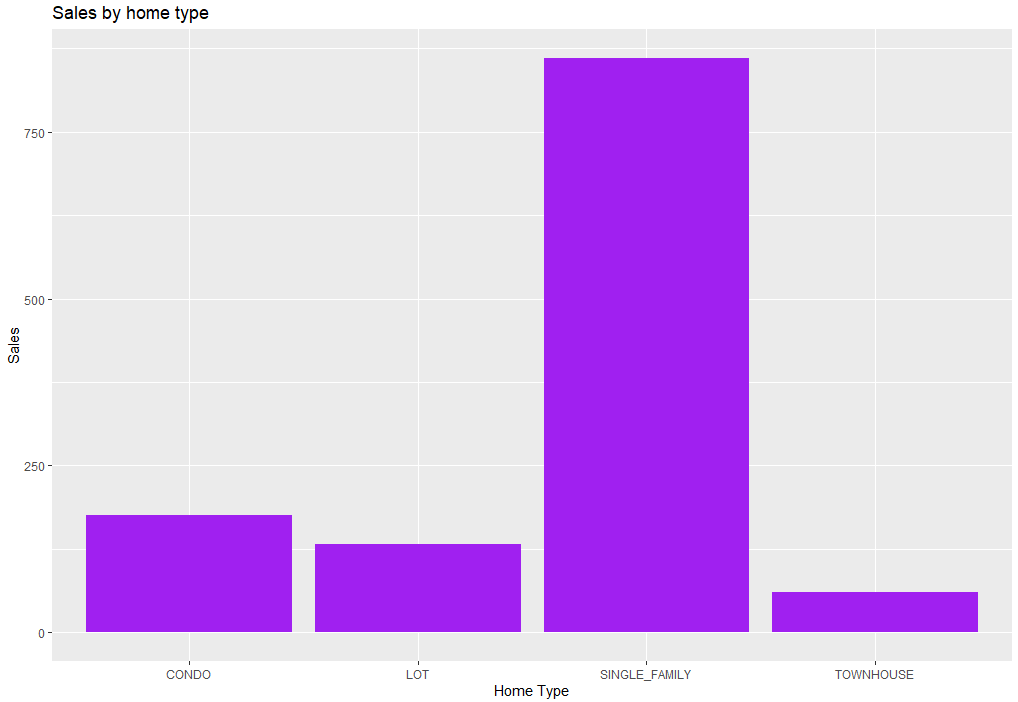
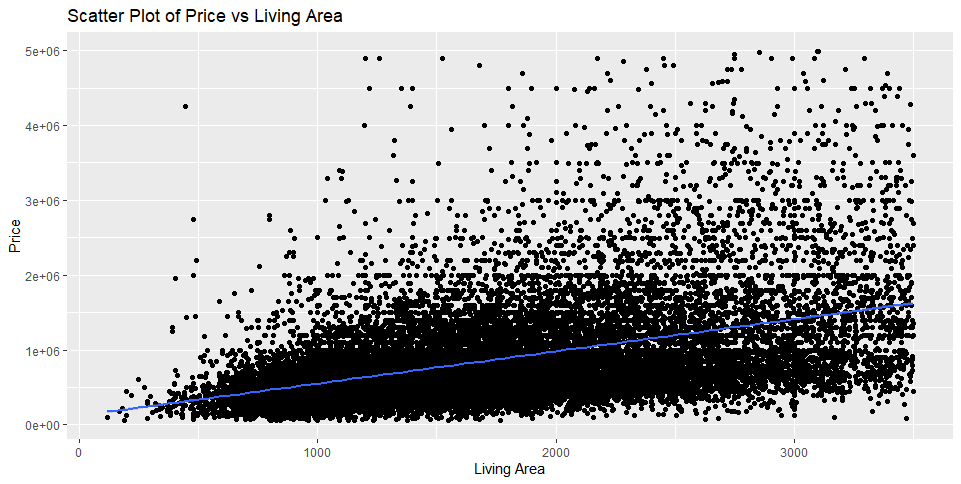


Figura 3 - Numărul de vânzări după tipul de locuinta

Graficul arată o reprezentare Scatter Plot (Figura 4) care compară prețul cu suprafața locuibilă. Se poate observa că există o corelație pozitivă între cele două variabile, deoarece pe măsură ce suprafața locuibilă crește, crește și prețul. Linia albastră indică regresia liniară a datelor, adică o linie care minimizează distanța dintre ea și punctele de date. Această linie arată tendința generală a datelor și poate fi folosită pentru a face predicții sau estimări despre prețul unei locuințe în funcție de suprafața sa locuibilă. Punctele negre reprezintă datele individuale, adică prețul și suprafața locuibilă a fiecărei locuințe din setul de date. Aceste puncte sunt răspândite pe tot parcursul graficului, indicând că există o variație semnificativă în date. Acest lucru înseamnă că alți factori pot influența prețul pe lângă suprafața locuibilă, cum ar fi locația, calitatea, tipul de locuință etc.

Figură 1 - Scatter plot price / livingArea

# Bibliografie

1. Dataset: [California Estate](https://www.kaggle.com/datasets/yellowj4acket/real-estate-california) - sursa setului de date utilizat în analiză, disponibil pe Kaggle. – accesat (10.09.2023)
2. R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL [https://www.R-project.org/](https://www.r-project.org/). – accesat (10.09.2023)
3. Wickham, H., și alții. (2019). *tidyverse: Easily Install and Load the 'Tidyverse'*. Disponibil la: [https://www.tidyverse.org](https://www.tidyverse.org/). – accesat(14.09.2023)
4. Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer. Disponibil la: [https://ggplot2.tidyverse.org](https://ggplot2.tidyverse.org/). – accesat(14.09.2023)
5. OpenIntro Statistics. (2019). [openintro: OpenIntro data sets and supplemental functions](https://cran.r-project.org/package=openintro). – accesat (04.11.2023)

# Materiale Suplimentare

Anexa 1. Variabilele setului de date

1. **ID Proprietate (id):** Reprezintă identificatorul unic al fiecărei proprietăți imobiliare.
2. **ID Stat (stateId):** Indică identificatorul statului în care se află proprietatea.
3. **ID Comitat (countyId):** Specifică identificatorul comitatului în care se găsește proprietatea.
4. **ID Oraș (cityId):** Reflectă identificatorul orașului în care este situată proprietatea.
5. **Țară (country):** Indică țara în care se află fiecare proprietate imobiliară.
6. **Dată Postare Anunț (datePostedString):** Furnizează informații despre data postării anunțului pentru proprietate.
7. **Deținută de Bancă (is\_bankOwned):** Indică dacă proprietatea este sau nu deținută de o bancă.
8. **Este la Licitație (is\_forAuction):** Specifică dacă proprietatea este sau nu disponibilă la licitație.
9. **Eveniment (event):** Reprezintă informații despre evenimentele asociate proprietății.
10. **Timp (time):** Indică marcajul temporal asociat fiecărui eveniment legat de proprietate.
11. **Preț (price):** Indică valoarea financiară a fiecărei proprietăți imobiliare.
12. **Preț pe Metru Pătrat (pricePerSquareFoot):** Reflectă prețul pe metru pătrat al proprietății.
13. **Oraș (city):** Specifică orașul în care se află fiecare proprietate.
14. **Stat (state):** Indică statul în care este situată proprietatea.
15. **An Construcție (yearBuilt):** Specifică anul în care a fost construită fiecare proprietate.
16. **Numele Proprietății (streetAddress):** Identifică adresa fiecărei proprietăți imobiliare analizate.
17. **Cod Poștal (zipcode):** Furnizează informații despre codul poștal al fiecărei proprietăți.
18. **Longitudine și Latitudine (longitude, latitude):** Reprezintă coordonatele geografice ale fiecărei proprietăți.
19. **Anul Construcției (yearBuilt):** Specifică anul în care a fost construită fiecare proprietate.
20. **Preț (price):** Indică valoarea financiară a fiecărei proprietăți imobiliare.
21. **Preț pe Metru Pătrat (pricePerSquareFoot):** Reflectă prețul pe metru pătrat al proprietății.
22. **Suprafața Utilă (livingArea):** Indică suprafața utilă a locuinței.
23. **Suprafața Lotului (lotAreaUnits):** Specifică unitatea de măsură a suprafeței lotului.
24. **Număr de Băi (bathrooms):** Reprezintă numărul total de băi din proprietate.
25. **Număr de Dormitoare (bedrooms):** Arată numărul total de dormitoare din proprietate.
26. **Suprafața Construită (buildingArea):** Indică suprafața totală construită a proprietății.
27. **Număr de Locuri de Parcare (parking):** Specifică numărul total de locuri de parcare disponibile.
28. **Număr de Spații de Garaj (garageSpaces):** Reprezintă numărul de spații de garaj disponibile.
29. **Disponibilitate Garaj (hasGarage):** Indică dacă proprietatea are sau nu un garaj.
30. **Existența Piscinei (pool):** Specifică dacă proprietatea are o piscină.
31. **Existența Spa-ului (spa):** Indică dacă proprietatea are un spațiu de relaxare cu jacuzzi.
32. **Stare Construcție Nouă (isNewConstruction):** Reflectă dacă proprietatea este sau nu în construcție nouă.
33. **Admitere Animale de Companie (hasPetsAllowed):** Specifică dacă proprietatea permite animale de companie.
34. **Tipul de Locuință (homeType):** Categorizează tipul de locuință, cum ar fi casă, apartament, etc.
35. **Comitatul (county):** Indică comitatul în care se află fiecare proprietate imobiliară.