**ACADEMIA DE STUDII ECONOMICE**

**FACULTATEA DE CIBERNETICA, STATISTICA SI INFORMATICA ECONOMICA**

**SPECIALIZAREA INFORMATICA ECONOMICA**

**Proiect**

**Dezvoltare software pentru analiza datelor**

**Studenti: Constantin Evelina Gabriela, Pîcălău Dumitru ,Padure Vlad**

**An 3, seria E**

**1. Obiectivul Proiectului**

Piața imobiliară este influențată de multe caracteristici ale proprietăților care lucrează împreună în moduri complexe. Prin acest proiect, încercăm să înțelegem mai bine cum aceste caracteristici influențează prețurile și cum sunt ele legate între ele.

Ne propunem să:

* Găsim cele mai importante pattern-uri în datele despre proprietăți
* Înțelegem cum diferite caracteristici ale caselor lucrează împreună pentru a influența prețul
* Oferim informații utile pentru toți cei implicați în piața imobiliară (cumpărători, vânzători, agenți imobiliari)
* Creăm o metodă de analiză care poate fi folosită și pentru alte seturi de date similare

**2. Sursa și Descrierea Datelor**

**2.1 Sursa Datelor**

Datele folosite în analiză vin de pe platforma Kaggle (<https://www.kaggle.com/datasets/reenapinto/housing-price-and-real-estate-2023>) și conțin informații despre proprietăți imobiliare din 2023. Am ales această sursă pentru că:

* Conține date recente despre piața imobiliară
* Include mai multe caracteristici importante ale proprietăților
* Are un număr suficient de mare de proprietăți pentru o analiză statistică relevantă

**2.2 Ce Informații Avem în Date?**

În setul nostru de date, pentru fiecare proprietate avem următoarele informații:

**Preț (Price)**:

* Reprezintă cât costă proprietatea în dolari americani
* E important pentru că e variabila pe care încercăm să o înțelegem
* În datele noastre, prețurile variază de la $141,900 până la $10,000,000
* Această variație mare ne permite să analizăm diferite segmente ale pieței

**Dormitoare (Beds)**:

* Numărul de camere destinate pentru dormit
* Variază de la 1 la 10 dormitoare în setul nostru
* Ne ajută să înțelegem capacitatea de locuit a proprietății
* E un factor important în decizia de cumpărare pentru multe familii

**Băi (Bath)**:

* Include atât băi complete cât și jumătăți de baie
* În date vedem variații de la 1 la 8 băi
* Numărul de băi poate indica nivelul de confort al proprietății
* Poate fi un indicator al modernității sau luxului unei case

**Suprafață (Sq.Ft)**:

* Măsoară spațiul util al proprietății în picioare pătrate
* În datele noastre variază de la 300 la 39,654 picioare pătrate
* Ne oferă o măsură obiectivă a mărimii proprietății
* E adesea unul din primii factori la care se uită cumpărătorii

**2.3 Ce Ne Spun Datele?**

Analizând datele, observăm câteva lucruri interesante:

**Varietatea Proprietăților**:

* Avem case de diferite mărimi și configurații
* Unele sunt mai compacte (307 picioare pătrate, 1 dormitoare)
* Altele sunt foarte spațioase (880 ,000picioare pătrate, 10 dormitoare)
* Această diversitate ne ajută să înțelegem diferite segmente ale pieței

**Relații între Caracteristici**:

* Casele mai mari tind să aibă mai multe dormitoare și băi
* Prețul nu crește întotdeauna proporțional cu mărimea
* Unele case mici pot fi mai scumpe dacă sunt într-o zonă bună

**3. Metode de Analiză**

**3.1 Analiza Componentelor Principale (PCA)**

**3.1.1 Ce este PCA și cum funcționează?**

Analiza Componentelor Principale (PCA – Principal Component Analysis) este o tehnică statistică utilizată pentru reducerea dimensionalității unui set de date, fără a pierde informația esențială. În esență, PCA încearcă să găsească un mod mai simplu de a privi datele, identificând structuri sau modele care ar putea să nu fie evidente la prima vedere. Aceasta se face prin transformarea datelor într-un nou sistem de coordonate, unde fiecare coordonată reprezintă o componentă principală.

**Cum funcționează PCA?**

1. **Pregătirea Datelor**:
   * Mai întâi, aducem toate măsurătorile la aceeași scară
   * Altfel, prețurile (care sunt în sute de mii) ar domina analiza față de numărul de dormitoare
   * Astfel, fiecare variabilă are o medie egală cu 0 și o abatere standard de 1, prevenind ca variabilele cu valori mari să domine analiza.
2. **Găsirea Pattern-urilor**:
   * PCA caută cele mai importante moduri în care variază datele
   * Prima componentă găsește cel mai puternic pattern
   * A doua componentă găsește următorul pattern important care e diferit de primul
   * Și așa mai departe
3. **Combinarea Informațiilor**:
   * Fiecare componentă combină informații din toate variabilele originale
   * De exemplu, prima componentă ar putea reprezenta "mărimea generală" a casei
   * A doua ar putea arăta cât de eficient e folosit spațiul

**3.1.2 De ce folosim PCA pentru datele noastre?**

PCA e utilă pentru analiza noastră din mai multe motive:

1. **Ne ajută să vedem imaginea de ansamblu**:
   * În loc să ne uităm separat la suprafață, dormitoare, băi și preț
   * PCA ne arată cum toate acestea lucrează împreună
   * Putem vedea pattern-uri care nu sunt evidente când privim fiecare caracteristică separat
2. **Simplifică analiza**:
   * Reducem patru variabile la două sau trei componente principale
   * Păstrăm majoritatea informației importante
   * E mai ușor să înțelegem și să vizualizăm datele
3. **Ne ajută să înțelegem ce e important**:
   * Vedem care caracteristici au cel mai mare impact
   * Înțelegem cum diferite aspecte ale unei case contribuie la preț
   * Putem identifica proprietăți neobișnuite sau speciale

**3.2 Ce ne spun rezultatele?**

**3.2.1 Analiza Fiecărei Componente Principale**

**A computer screen shot of a black screen

Description automatically generated**

**Prima Componentă (PC1) - "Mărimea Generală"**:

* Explică 68.95% din variația în date
* Toate caracteristicile contribuie pozitiv:
  + Băile au cea mai mare contribuție (0.541)
  + Urmează suprafața (0.497)
  + Apoi dormitoarele (0.489)
  + Și prețul (0.472)
* Ne spune că proprietățile mai mari tind să aibă mai multe băi, mai multe dormitoare și prețuri mai mari

**A Doua Componentă (PC2) - "Eficiența Spațiului"**:

* Explică încă 16.26% din variație
* Are contribuții interesante:
  + Pozitivă mare de la dormitoare (0.636)
  + Negativă mare de la preț (-0.602)
  + Negativă de la suprafață (-0.380)
* Ne arată proprietăți care au multe dormitoare dar nu sunt neapărat scumpe
* Poate indica eficiența cu care e folosit spațiul

**A Treia Componentă (PC3) - "Raportul Preț-Suprafață"**:

* Explică 9.36% din variație
* Contribuții principale:
  + Foarte pozitivă de la suprafață (0.773)
  + Foarte negativă de la preț (-0.622)
* Ne ajută să identificăm case care sunt mari dar relativ ieftine
* Util pentru găsirea oportunităților de investiție

**A Patra Componentă (PC4) - "Raportul Băi-Dormitoare"**:

* Explică ultimele 5.44% din variație
* Contribuții importante:
  + Pozitivă mare de la băi (0.781)
  + Negativă de la dormitoare (-0.592)
* Ne arată case cu multe băi raportat la numărul de dormitoare
* Poate indica nivelul de lux al proprietății

**3.2.2 Ce ne spun vizualizările?**

**Matricea de Corelație**:

* Ne arată cum sunt legate caracteristicile între ele
* Vedem că suprafața e strâns legată de numărul de dormitoare
* Prețul e influențat de toate celelalte caracteristici
* Ne ajută să înțelegem relațiile de bază din date

A screenshot of a graph

Description automatically generated

**Graficul Varianței Cumulative**:

* Ne arată cât de bine explicăm datele
* Primele două componente explică 85.20% din informație
* Asta înseamnă că putem folosi doar două componente pentru majoritatea analizelor
* E o simplificare bună a datelor originale

A graph with a line

Description automatically generated

**Heatmap-ul PCA**:

* Ne arată vizual contribuțiile la fiecare componentă
* Culorile intense indică contribuții mari
* Ne ajută să vedem rapid ce e important în fiecare componentă
* Face mai ușoară interpretarea componentelor

A screenshot of a graph

Description automatically generated

**4. Ce am învățat din această analiză?**

**4.1 Descoperiri Principale**

1. **Despre structura pieței**:
   * Există o legătură puternică între mărimea casei și preț
   * Dar relația nu e perfectă - alte factori sunt și ei importanți
   * Numărul de băi are o influență surprinzător de mare asupra prețului
2. **Despre eficiența folosirii spațiului**:
   * Unele case folosesc spațiul mai eficient decât altele
   * Poți avea multe dormitoare într-un spațiu relativ mic
   * Acest lucru nu reduce neapărat prețul casei
3. **Despre oportunități**:
   * Există case mari care sunt relativ ieftine
   * Unele proprietăți au caracteristici de lux (multe băi) dar prețuri moderate
   * Aceste informații pot fi valoroase pentru investitori

**3.2.3 Cum putem folosi aceste informații?**

**Pentru Cumpărători**:

* Pot înțelege mai bine ce influențează prețurile
* Pot identifica case care oferă valoare bună pentru bani
* Pot face comparații mai informate între proprietăți

**Pentru Vânzători și Agenți**:

* Pot evalua mai bine proprietățile
* Înțeleg ce caracteristici sunt mai valoroase
* Pot explica mai bine prețurile cerute

**Pentru Investitori**:

* Pot identifica proprietăți sub sau supraevaluate
* Înțeleg mai bine diferitele segmente ale pieței
* Pot face alegeri mai informate

Această analiză ne-a ajutat să înțelegem mai bine cum funcționează piața imobiliară și ce face ca o proprietate să fie valoroasă. Informațiile acestea pot ajuta pe oricine e implicat în tranzacții imobiliare să ia decizii mai bune și mai informate.

**3.3 Analiza Clusterelor Ierarhice**

**3.3.1 Fundamentele Analizei de Clustering**

Analiza de clustering ierarhic este o metodă de segmentare a datelor care organizează observațiile în grupuri (clustere) bazate pe similaritatea dintre caracteristicile lor. Această tehnică este deosebit de utilă în analiza pieței imobiliare deoarece:

* Permite identificarea naturală a segmentelor de piață
* Nu necesită specificarea în avans a numărului exact de clustere
* Oferă o structură ierarhică care poate fi analizată la diferite niveluri de granularitate

**3.3.2 Rezultatele Clusterizării**

**A screenshot of a computer

Description automatically generated**

În urma analizei, datele s-au grupat natural în trei segmente distincte ale pieței imobiliare:

**Clusterul 0 - Segment Premium Mediu**:

* Număr de proprietăți: 1,854 unități
* Preț mediu: $935,579
* Caracteristici:
  + 3.95 dormitoare în medie
  + 3.12 băi în medie
  + Suprafață medie de 1,974.21 sq.ft
* Range de preț: $299,900 - $4,000,000

**Clusterul 1 - Segment Accesibil**:

* Număr de proprietăți: 1,465 unități
* Preț mediu: $409,609
* Caracteristici:
  + 2.04 dormitoare în medie
  + 1.60 băi în medie
  + Suprafață medie de 919.03 sq.ft
* Range de preț: $141,900 - $1,200,000

**Clusterul 2 - Segment Luxury**:

* Număr de proprietăți: 41 unități
* Preț mediu: $4,526,206
* Caracteristici:
  + 4.73 dormitoare în medie
  + 4.79 băi în medie
  + Suprafață medie de 5,635.54 sq.ft
* Range de preț: $1,695,000 - $10,000,000

**3.3.3 Interpretarea Rezultatelor**

**Segmentarea Naturală a Pieței**:

* Analiza a identificat trei segmente clare ale pieței, fiecare cu caracteristici distincte
* Distribuția proprietăților între clustere reflectă structura piramidală tipică a pieței imobiliare
* Există o corelație clară între dimensiunile proprietăților și segmentul de preț

**Caracteristici Distinctive ale Segmentelor**:

1. Segmentul Accesibil:
   * Reprezintă aproximativ 43% din piață
   * Proprietăți mai mici, orientate către familii mici sau cupluri
   * Raport băi/dormitoare aproape de 0.8
2. Segmentul Premium Mediu:
   * Reprezintă aproximativ 55% din piață
   * Proprietăți medii spre mari, potrivite pentru familii
   * Raport băi/dormitoare de aproximativ 0.79
3. Segmentul Luxury:
   * Reprezintă aproximativ 2% din piață
   * Proprietăți foarte mari cu facilități extensive
   * Raport băi/dormitoare de peste 1, indicând un nivel ridicat de lux

**Implicații pentru Piață**:

* Există o separare clară între segmentele de piață în termeni de preț și caracteristici
* Saltul de la segmentul premium mediu la luxury este mult mai mare decât cel de la accesibil la premium
* Numărul mic de proprietăți din segmentul luxury (41) indică exclusivitatea acestui segment

**3.3.4 Corelații cu Analiza PCA**

Rezultatele clusterizării se aliniază cu observațiile din analiza PCA:

* Prima componentă principală (mărimea generală) separă clar segmentul luxury de celelalte
* A doua componentă (eficiența spațiului) diferențiază între segmentele accesibil și premium mediu
* Raportul băi/dormitoare identificat în PC4 este relevant în special pentru segmentul luxury

**4. Concluzie**

Această analiză a pieței imobiliare a evidențiat factori și pattern-uri esențiale care influențează prețurile proprietăților și structura pieței. Utilizând metode precum PCA și analiza clusterelor ierarhice, am obținut o înțelegere detaliată a relațiilor dintre caracteristicile caselor și prețurile acestora. Iată principalele concluzii:

**Descoperiri Cheie:**

1. **Structura Pieței:**
   * Piața imobiliară poate fi segmentată în trei categorii distincte: accesibil, premium mediu și luxury.
   * Fiecare segment are caracteristici proprii în termeni de preț, dimensiuni și raportul băi/dormitoare.
   * Segmentul luxury, deși foarte restrâns, are caracteristici clar delimitate, indicând un nivel înalt de exclusivitate.
2. **Factorii care Influențează Prețul:**
   * Mărimea generală (PC1) este principalul factor de diferențiere a proprietăților.
   * Eficiența utilizării spațiului (PC2) oferă informații despre modul în care casele optimizează caracteristicile lor pentru a oferi valoare.
   * Raportul preț/suprafață (PC3) și raportul băi/dormitoare (PC4) identifică proprietăți unice, care pot reprezenta oportunități de investiție.
3. **Valoare pentru Actorii Pieței:**
   * Cumpărătorii pot folosi aceste informații pentru a identifica proprietăți subevaluate sau cu raport bun între preț și caracteristici.
   * Vânzătorii și agenții imobiliari pot înțelege mai bine cum să-și poziționeze proprietățile pe piață și să justifice prețurile cerute.
   * Investitorii pot identifica segmente de piață și proprietăți care oferă cel mai bun potențial de rentabilitate.
4. **Utilitatea Metodelor:**
   * PCA simplifică analiza datelor complexe, păstrând informațiile esențiale și oferind o perspectivă de ansamblu asupra pieței.
   * Analiza clusterelor oferă o segmentare naturală a pieței, fiind utilă pentru identificarea grupurilor de proprietăți similare.

**Perspective Viitoare:**

Analiza poate fi extinsă pentru a include alți factori relevanți, cum ar fi localizarea, facilitățile comunitare și tendințele economice. Integrarea unor metode avansate de machine learning ar putea oferi predicții mai precise despre prețurile viitoare și preferințele cumpărătorilor.

În concluzie, aceste date și analize reprezintă un instrument valoros pentru toți actorii implicați în piața imobiliară, ajutând la luarea unor decizii mai bine fundamentate și mai strategice.