

Universitatea Babes-Bolyai  
Facultatea de Științe Economice și Gestiunea Afacerilor

*Predicția prețului unei locuințe*

Hurmuz Alicia-Maria (alicia.hurmuz@stud.ubbcluj.ro)  
Iaroi Renata-Timea (renata.iaroi@stud.ubbcluj.ro)  
Informatică Economică  
An 3

## Introducere

În prezent, numărul persoanelor care vor să își cumpere o locuință proprie, crește semnificativ, în special în rândul tinerilor. Locuința este cel mai important factor din viața unei persoane deoarece influențează pozitiv starea de bine, ne oferă siguranță și protecție. De aceea, trebuie să înțelegem factorii care influențează tendința de creștere a prețurilor, prețul fiind important în decizia cumpărării unui imobil.

Am ales să ne concentrăm asupra întrebării:

*Cum influențează numărul dormitoarelor, suprafața în metri pătrați ( $m^2$ ), numărul băilor, etajele și anul construcției imobilului prețul de cumpărare?*

Această întrebare de cercetare are ca și scop oferirea de informații valoroase despre cum factorii de mai sus, influențează stabilirea prețului. Mai mult, informațiile sunt un real interes în achiziționarea unui imobil, pentru a înțelege mai bine piața imobiliarelor și stabilirea deciziei finale în funcție de dorințele și necesitățile viitorilor cumpărători.

Cât de importanți sunt factorii studiați?

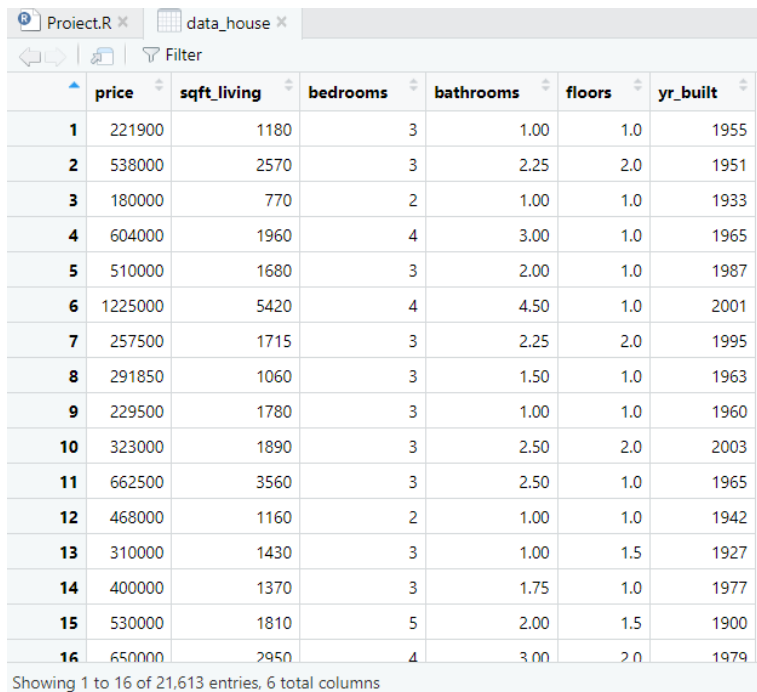
Numărul dormitoarelor este foarte evident, mai multe dormitoare înseamnă preț mai mare. Confortul și spațiul suplimentar ridică prețul casei. Suprafața în metri pătrați( $m^2$ ) este direct proporțională cu prețul. Numărul băilor cresc valoarea locuinței, crește și confortul pentru familii mari. Etajele, în special pentru apartamentele care sunt la nivele superioare prețul poate fi mult mai mare iar anul construcției are un impact semnificativ pentru că tehnologiile de construcție apărute în ultimii ani și modernizarea materialelor de construcție pot influența creșterea prețului.

## Setul de date

Setul pe care l-am ales este <https://www.kaggle.com/datasets/soylevbeytullah/house-prices-dataset>. Acesta conține 21 de coloane, dar noi am lăsat doar variabilele relevante pentru întrebarea de cercetare aleasă, acestea fiind:

- **Price:** prețul casei,
- **Sqft\_living:** suprafața locuibilă a casei,
- **Bedrooms:** numărul de dormitoare,
- **Bathrooms:** numărul de băi,
- **Floors:** numărul de etaje,
- **Yr\_built:** anul când a fost construită construcția.

Am adus câteva modificări setului de date, precum eliminarea valorilor nule și apoi selectarea doar a coloanelor care sunt de interes pentru întrebarea noastră de cercetare. Aceste lucruri le-am făcut în R Studio.



	price	sqft_living	bedrooms	bathrooms	floors	yr_built
1	221900	1180	3	1.00	1.0	1955
2	538000	2570	3	2.25	2.0	1951
3	180000	770	2	1.00	1.0	1933
4	604000	1960	4	3.00	1.0	1965
5	510000	1680	3	2.00	1.0	1987
6	1225000	5420	4	4.50	1.0	2001
7	257500	1715	3	2.25	2.0	1995
8	291850	1060	3	1.50	1.0	1963
9	229500	1780	3	1.00	1.0	1960
10	323000	1890	3	2.50	2.0	2003
11	662500	3560	3	2.50	1.0	1965
12	468000	1160	2	1.00	1.0	1942
13	310000	1430	3	1.00	1.5	1927
14	400000	1370	3	1.75	1.0	1977
15	530000	1810	5	2.00	1.5	1900
16	650000	2950	4	3.00	2.0	1979

Showing 1 to 16 of 21,613 entries, 6 total columns

## Rezultate și discuții

Pentru a răspunde la întrebarea „Cum influențează numărul dormitoarelor, suprafața în metri pătrați ( $m^2$ ), numărul băilor, etajele și anul construcției imobilului prețul de cumpărare? vom folosi, în primul rând, **regresia liniară**.

Vom alege variabila dependentă „price” pentru toate regresiile pe care le vom efectua în continuare.

Prima regresie liniară are variabila independentă „sqft\_living”.

```
Call:
lm(formula = price ~ sqft_living, data = data_house)

Residuals:
    Min       1Q   Median       3Q      Max
-1476062  -147486   -24043   106182   4362067

Coefficients:
            Estimate Std. Error
(Intercept) -43580.743    4402.690
sqft_living    280.624      1.936
            t value Pr(>|t|)
(Intercept)  -9.899  <2e-16 ***
sqft_living  144.920  <2e-16 ***
---
Signif. codes:
  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 261500 on 21611 degrees of freedom
Multiple R-squared:  0.4929,    Adjusted R-squared:  0.4928 
F-statistic: 2.1e+04 on 1 and 21611 DF,  p-value: < 2.2e-16
```

Formula:  $price \approx \beta_0 + \beta_1 \times sqft\_living$

Interceptul( $\beta_0$ ) sugerează că valoarea prețului atunci când sqft\_living ar fi 0.

Coeficientul sqft\_living( $\beta_1$ ) de 280,624 sugerează cu câte unități monetare crește prețul casei, în vreme ce suprafața locuibilă crește.

Std. Error ilustrează cu cât diferă interceptul de valoarea reală, adică cu 4402,69 de unități monetare. În cazul coeficientului sqft\_living, eroarea standard e de doar 1,936 unități monetare, fiind o eroare mică. Calculând aceste valori în valoare relativă pentru o concluzie mai clară:

$(4402,690/43580,743) \times 100 = 10,10\%$  și  $(1,936/280,634) \times 100 = 0,69\%$ , sugerând că eroarea pentru coeficientul sqft\_living e mai mică.

T statistic ilustrează cu cât se diferențiază parametrul calculat  $\beta$  de 0.

P value arată probabilitatea ca între predictor și variabila independentă poate exista o asociere datorită șansei. Un p value mare indică faptul că asocierea se datorează șansei. În caz contrar, un p value mic înseamnă ca asocierea se datorează factorului. În cazul modelului nostru, observăm că

atât interceptul cât și sqft\_living au o valorile p foarte mici ( $<2e-16$ ), sugerând că sunt foarte semnificative în ceea ce constă prețul locuinței.

RSE-ul din cadrul acestei regresii, care e 261500, indică cu cât sqft\_living se distanșează de la dreapta regresiei, adică valorile reale(ale prețului), de valorile modelului(pe baza sqft\_living). Această valoare determină R-squared care este 0,4929, sugerând că modelul nostru explică 49,29% din fluctuațiile prețurilor locuințelor în funcție de dimensiunea imobilului.

```
> confint(mod_price_sqft_living)
              2.5 %      97.5 %
(Intercept) -52210.3396 -34951.1466
sqft_living   276.8281    284.4191
> |
```

Intervalele de încredere arată că interceptul se află în intervalul [-52210,34, -34951,15], iar coeficientul sqft\_living se află între [-276,83 , 284,42].

A doua regresie liniară va prelua numărul de dormitoare („bedrooms”) ca variabilă independentă.

```
Call:
lm(formula = price ~ bedrooms, data = data_house)

Residuals:
    Min       1Q   Median       3Q      Max
-3506435 -203235  -66667   105049  6839901

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   129802      8932    14.53  <2e-16 ***
bedrooms      121716      2554    47.65  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 349200 on 21611 degrees of freedom
Multiple R-squared:  0.09508, Adjusted R-squared:  0.09504
F-statistic: 2271 on 1 and 21611 DF, p-value: < 2.2e-16
```

Formula:  $price \approx \beta_0 + \beta_1 \times bedrooms$ .

Interceptul( $\beta_0$ ) sugerează că valoarea prețului atunci când bedrooms ar fi 0.

Coeficientul bedrooms( $\beta_1$ ) de 1299802 sugerează cu câte unități monetare crește prețul imobilului, în timp ce numărul de dormitoare ale unei locuințe crește..

Std. Error prezintă cu cât diferă interceptul de valoarea reală, adică cu 8932 de unități monetare. În cazul coeficientului bedrooms, eroarea standard e de 2554. Calculând aceste valori în valoare relativă pentru o concluzie mai clară

$(8932/129802) \times 100 = 6,88\%$  și  $(2554/121716) \times 100 = 2,10\%$ , sugerând că eroarea pentru coeficientul bedrooms e mai mică.

T statistic cu cât se diferențiază parametrul calculat  $\beta$  de 0. Ambele valori t arată că sunt diferite de 0.

În cazul modelului nostru, observăm că atât interceptul cât și coeficientul bedrooms au o valorile p foarte mici ( $< 2e-16$ ), sugerând că sunt foarte semnificative în ceea ce constă prețul locuinței.

RSE-ul din cadrul acestei regresii, care e 349200, indică diferența dintre valorile reale (ale prețului), de valorile modelului (pe baza bedrooms). Această valoare determină R-squared care este 0.09508, sugerând că modelul nostru explică 9,508% din fluctuațiile prețurilor locuințelor în funcție de numărul de dormitoare, nefiind foarte sugestiv.

```
> confint(mod_price_bedrooms)
                2.5 %    97.5 %
(Intercept) 112295.2 147309.5
bedrooms    116709.5 126722.8
```

Intervalele de încredere arată că interceptul se află în intervalul [112295.2, 147309.5], iar coeficientul bedrooms se află între [116709.5, 126722.8].

A treia regresie are ca variabilă dependentă numărul de băi („bathrooms”).

```
Call:
lm(formula = price ~ bathrooms, data = data_house)

Residuals:
    Min       1Q   Median       3Q      Max
-1438157 -184525  -41525   113220  5925322

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    10708      6211    1.724   0.0847 .
bathrooms    250327      2760   90.714  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 312400 on 21611 degrees of freedom
Multiple R-squared:  0.2758,    Adjusted R-squared:  0.2757
F-statistic: 8229 on 1 and 21611 DF,  p-value: < 2.2e-16
```

Formula:  $price \approx \beta_0 + \beta_1 \times bathrooms$ .

Interceptul ( $\beta_0$ ) sugerează că valoarea prețului atunci când bathrooms ar fi 0.

Coeficientul bathrooms ( $\beta_1$ ) de 250327 sugerează cu câte unități monetare crește prețul casei, în timp ce numărul de băi ale unei locuințe crește.

Valoarea t pentru intercept e 1,724 și nu este semnificativă. În schimb, valoarea t a coeficientului bathrooms este 90,714, fiind foarte semnificativă.

Std. Error prezintă cu cât diferă interceptul de valoarea reală, adică cu 6211 de unități monetare. În cazul coeficientului bathrooms, eroarea standard e de 2760. Calculând aceste valori în valoare relativă pentru o concluzie mai clară:

$(6211/10708) \times 100 = 58,02\%$  și  $(2760/250327) \times 100 = 1,01\%$ , sugerând că eroarea pentru coeficientul bathrooms e cu mult mai mică.

În cazul modelului nostru, observăm că coeficientul bathrooms are valoarea p foarte mică, sugerând că este foarte semnificativ în ceea ce constă prețul locuinței. În schimb, valoarea p a interceptului nu este semnificativă, aceasta fiind 0,0847.

RSE-ul din cadrul acestei regresii, care e 312400, indică diferența dintre valorile reale (ale prețului), de valorile modelului (pe baza bathrooms). Această valoare determină R-squared care este 0.02758, sugerând că modelul nostru explică 27,58% din fluctuațiile prețurilor locuințelor în funcție de numărul de băi, fiind mai sugestiv decât modelul anterior (pe baza coeficientului bedrooms).

```
> confint(mod_price_bathrooms)
              2.5 %      97.5 %
(Intercept) -1465.06  22881.68
bathrooms    244917.64 255735.39
```

Intervalele de încredere arată că interceptul se află în intervalul [-1465.06, 22881.68], iar coeficientul bedrooms se află între [244917.64, 255735.39].

Continuăm cu alegerea variabilei dependente „floors”(etaje).

```
Call:
lm(formula = price ~ floors, data = data_house)

Residuals:
    Min       1Q   Median       3Q      Max
-597965 -203837  -73787   103213  6984329

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   279199      7102    39.31  <2e-16 ***
floors        174589      4470    39.06  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 354800 on 21611 degrees of freedom
Multiple R-squared:  0.06594,    Adjusted R-squared:  0.0659
F-statistic: 1526 on 1 and 21611 DF,  p-value: < 2.2e-16
```

Formula:  $price \approx \beta_0 + \beta_1 \times floors$ .

Interceptul( $\beta_0$ ) sugerează că valoarea prețului atunci când numărul de etaje ar fi 0.

Coeficientul floors ( $\beta_1$ ) de 174589 sugerează cu câte unități monetare crește prețul casei, în timp ce numărul de etaje ale unei locuințe crește.

Valorile t pentru intercept și pentru coeficientul floors sunt 39,31 și 39,06, fiind foarte semnificative.

Std. Error prezintă cu cât diferă interceptul de valoarea reală, adică cu 7102 de unități monetare. În cazul coeficientului floors, eroarea standard e de 4470. Calculând aceste valori în valoare relativă pentru o concluzie mai clară:

$(7102/279199) \times 100 = 2,57\%$  și  $(4470/279199) \times 100 = 2,56\%$ , sugerând că erorile sunt asemănătoare.

În cazul modelului nostru, observăm că atât interceptul cât și coeficientul floors au o valorile p foarte mici, sugerând că sunt foarte semnificative în ceea ce constă prețul locuinței.

RSE-ul din cadrul acestei regresii, care e 354800, indică diferența dintre valorile reale (ale prețului), de valorile modelului (pe baza floors). Această valoare determină R-squared care este 0.06594, sugerând că modelul nostru explică 6,59% din fluctuațiile prețurilor locuințelor în funcție de numărul de etaje, nefiind foarte sugestiv.

```
> confint(mod_price_floors)
                2.5 %    97.5 %
(Intercept) 265278.4 293118.7
floors      165827.8 183349.8
```

Intervalele de încredere arată că interceptul se află în intervalul [265278.4, 293118.7], iar coeficientul floors se află între [165827.2, 183350.8].

Vom continua prin alegerea variabilei dependente „yr\_built” (anul construcției).

```
Call:
lm(formula = price ~ yr_built, data = data_house)

Residuals:
    Min       1Q   Median       3Q      Max
-461709 -221337  -87006   104064  7201095

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -790477.9    167350.2  -4.723 2.33e-06 ***
yr_built      675.1         84.9    7.952 1.93e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 366600 on 21611 degrees of freedom
Multiple R-squared:  0.002917, Adjusted R-squared:  0.002871
F-statistic: 63.23 on 1 and 21611 DF, p-value: 1.93e-15
```

Formula:  $price \approx \beta_0 + \beta_1 \times yr\_built$ .



Coeficientul `yr_built` ( $\beta_1$ ) de 675,1 sugerează cu câte unități monetare crește prețul casei, în timp ce anul de construcție ale unei locuințe crește.

Valorile  $t$  pentru `intercept` și pentru coeficientul `floors` -7,723 și 7,952, fiind semnificative și diferite de 0.

Std. Error prezintă cu cât diferă interceptul de valoarea reală, adică cu 167350,2 de unități monetare. În cazul coeficientului `yr_built`, eroarea standard e de 84,9. Calculând aceste valori în valoare relativă pentru o concluzie mai clară:

$(167350.2/675,1) \times 100 = 21,17\%$  și  $(84,9/675,1) \times 100 = 12,57\%$ , sugerând că erorile standard pentru `yr_built` e mai mică.

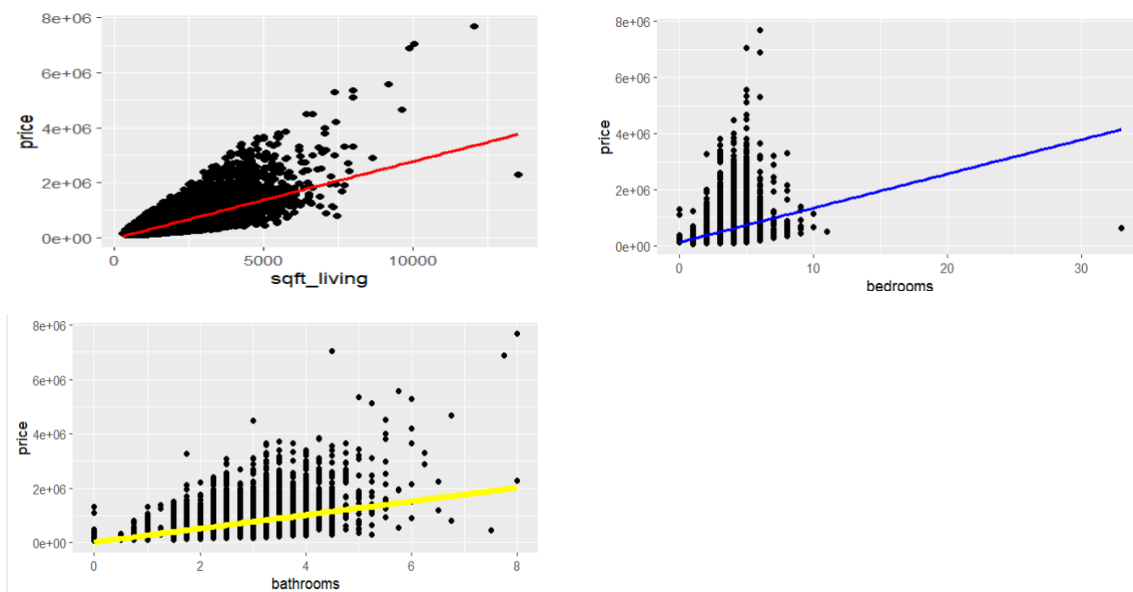
În cazul modelului nostru, observăm că atât interceptul cât și coeficientul `floors` au o valorile  $p$  destul de semnificative.

RSE-ul din cadrul acestei regresii, care e 366600, indică diferența dintre valorile reale (ale prețului), de valorile modelului (pe baza anului construcției). Această valoare determină  $R$ -squared care este 0.002917, sugerând că modelul nostru explică doar 0,2917% din fluctuațiile prețurilor locuințelor în funcție de anul construcției, nefiind sugestiv.

```
> confint(mod_price_yr_built)
                2.5 %      97.5 %
(Intercept) -1118496.6760 -462459.0697
yr_built      508.6662      841.4734
```

Intervalele de încredere arată că interceptul se află în intervalul  $[-1118496,6760, -462459,0697]$ , iar coeficientul `yr_built` se află între  $[508,6662, 841,4734]$ .

Cele mai relevante grafice ale regresiilor sunt



Se poate remarca în mod cert faptul că cel mai puternic predictor este `sqft_living`, adică dimensiunea locuinței. Apoi, al doilea cel mai relevant grafic este relația dintre preț și numărul de băi, remarcându-se o relație pozitivă și clară. Nu în ultimul rând, avem relația dintre preț și numărul de camere, având o relație vizibilă, dar nu așa puternică ca celelalte două.

În continuare, vom adauga toate cele 5 variabile independente în regresie și vom analiza rezultatul.

```
Call:
lm(formula = price ~ sqft_living + bedrooms + bathrooms + floors +
    yr_built, data = data_house)

Residuals:
    Min       1Q   Median       3Q      Max
-1838977 -130032  -15629   97819 3952758

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 6588743.65  134785.75   48.88  <2e-16 ***
sqft_living    299.29     2.94  101.81  <2e-16 ***
bedrooms     -67222.30   2238.02  -30.04  <2e-16 ***
bathrooms     68590.53   3834.14   17.89  <2e-16 ***
floors        55863.67   3766.09   14.83  <2e-16 ***
yr_built     -3385.62    69.95   -48.40  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 244900 on 21607 degrees of freedom
Multiple R-squared:  0.5551,    Adjusted R-squared:  0.555
F-statistic: 5393 on 5 and 21607 DF,  p-value: < 2.2e-16
```

### Interpretari:

- `Sqft_living`- Pentru un nivel fix al al celorlalte variabile, coeficientul este de 299.99, însemnând prețul unei locuințe va crește cu 299,29 de unități monetare, pentru fiecare creșterea cu o unitate a suprafeței locuinței.
- `Bathrooms`- Pentru un nivel fix al al celorlalte variabile, coeficientul este de 68590.53, însemnând prețul unei locuințe va crește cu 68590.53 de unități monetare, pentru fiecare creștere cu o unitate a numărului de băi.
- `Floors`- Pentru un nivel fix al al celorlalte variabile, coeficientul este de 55863.67, însemnând prețul unei locuințe va crește cu 55863.67 de unități monetare, pentru fiecare creștere cu o unitate a numărului de etaje ale locuinței.
- `Yr_built`- Pentru un nivel fix al al celorlalte variabile, coeficientul este de -3385.62, însemnând prețul unei locuințe va scădea cu -3385.62 de unități monetare, pentru fiecare an care trece de la construcția locuinței.
- `Bedrooms`- Pentru un nivel fix al al celorlalte variabile, coeficientul este de -67222.30, însemnând prețul unei locuințe va scădea cu 67222,30 de unități monetare, pentru fiecare creștere cu o unitate a numărului de dormitoare.

În final, putem trage următoarele concluzii:

1. Modelul se potrivește în proporție de 55,51% pe datele disponibile, lucru rezultat din valoarea lui R-squared.
2. Datorită valorilor p foarte mici ( $< 2e-16$ ), constatăm că variabilele sunt relevante pentru a prezice variabila dependentă price.
3. În explicarea variabilei dependente price, toți predictorii au o valoare explicativă bună, remarcându-se cel mai mult și pozitiv variabilele sqft\_living, floors și barhrooms.

Până acum am construit 6 modele, 5 regresii simple și una multiplă.

Așadar, putem observa că cel mai bun model până acum este cel de regresie multiplă, având cel mai mare R-squared, cea mai mică valoare RSE și un F-statistic de 5393, valoare îndepărtată de 1, indicând o relație puternică.

Vom continua raportul prin realizarea predicțiilor pentru un set de date nou, utilizând modelul de regresie antrenat anterior. Noul set de date conține următoarele date:

- sqft\_living = 2500,
- bedrooms = 4,
- bathrooms = 2.5,
- floors = 2,
- yr\_built = 2005.

```
> predict(mod_price_all, newdata = new_house, interval = "confidence")
      fit      lwr      upr
1 563104.7 557400.3 568809.1
> predict(mod_price_all, newdata = new_house, interval = "prediction")
      fit      lwr      upr
1 563104.7 83065.09 1043144
```

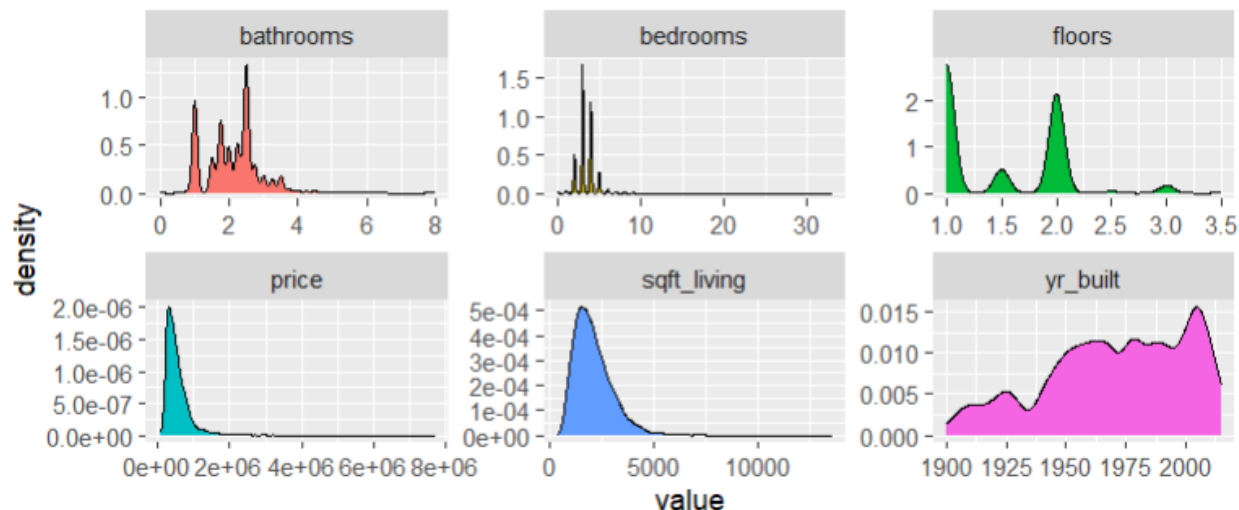
Intervalul de încredere se situează între limita inferioară(lwr) de 557500,3 și limita superioară (upr) de 568809,1. Se estimează pe baza acestui interval de încredere (de 95%), prețul mediu al locuinței pe acest set nou de date, va fi 563104,7.

Intervalul de predicție se situează între limita inferioară(lwr) de 563104,7 și limita superioară (upr) de 83065,09. Se estimează pe baza acestui interval de predicție, cu o încredere de de 95%, prețul unei locuințe individuale pe acest set nou de date, va fi 563104,7. Aici, intervalul este mai larg deoarece măsoară incertitudinea unei valori individuale asupra prețului.

Mai mult, valoarea RMSE pentru această regresie este 247177.7 sugerând cu cât diferă valorile reale de predicțiile făcute de model, aceasta fiind o valoare medie.

Al doilea tip de model ce urmează a fi prezentat este arborele de decizie.

Graficul densităților prezintă distribuția valorilor variabilelor numerice din setul de date, în datele de antrenament.



Se pot observa diferite aspecte, precum faptul că cele mai frecvente locuințe au una sau 2 băi. Totodată, numărul de dormitoare variază între 2 și 4. Majoritatea locuințelor au 1 sau 2 etaje, dar se observă și locuințe cu 3 etaje. Graficul distribuției prețurilor arată că majoritatea locuințelor au prețuri mai mici. De asemenea, putem vedea că cele mai multe construcții de locuințe au fost construite între 1950 și 2000.

Pentru început am împărțit setul de date în două, unul set pentru antrenament(70%) și unul pentru test (30%), apoi am creat arborele de antrenament.

n= 15129

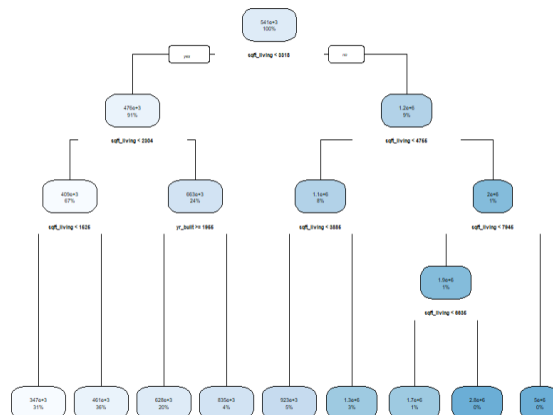
node), split, n, deviance, yval  
\* denotes terminal node

```

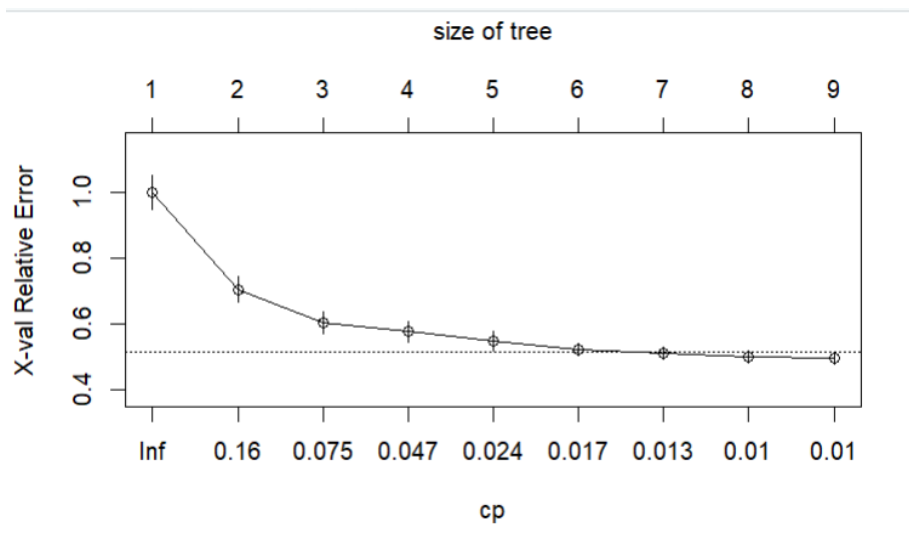
1) root 15129 2.046130e+15 540934.2
2) sqft_living< 3318 13724 7.452131e+14 475764.5
4) sqft_living< 2304 10111 2.969374e+14 408818.9
8) sqft_living< 1525 4622 8.603097e+13 347251.4 *
9) sqft_living>=1525 5489 1.786339e+14 460661.6 *
5) sqft_living>=2304 3613 2.761476e+14 663112.3
10) yr_built>=1954.5 3006 1.893344e+14 628328.9 *
11) yr_built< 1954.5 607 6.516540e+13 835367.9 *
3) sqft_living>=3318 1405 6.732852e+14 1177509.0
6) sqft_living< 4755 1219 3.055553e+14 1056059.0
12) sqft_living< 3885 759 1.089402e+14 922794.4 *
13) sqft_living>=3885 460 1.608946e+14 1275946.0 *
7) sqft_living>=4755 186 2.319101e+14 1973465.0
14) sqft_living< 7945 179 1.340012e+14 1853845.0
28) sqft_living< 6635 157 8.537230e+13 1727058.0 *
29) sqft_living>=6635 22 2.809460e+13 2758643.0 *
15) sqft_living>=7945 7 2.985104e+13 5032329.0 *

```

> |

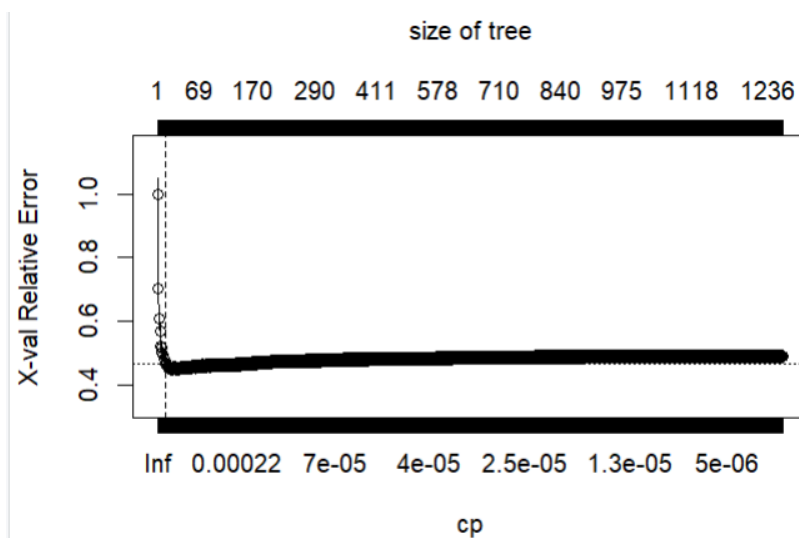


Setul de date este împărțit în mai multe subseturi de dimensiuni mai mici. Avem nodul de început, nodul rădăcina= 15129 care conține instanțele. Urmează a se împărți setul de date în două, pe baza variabilei criteriului `sqft_living<3318` pentru minimizarea SSE(devierea). Nodurile care conțin predicția finală sunt prezente în primul grafic (text) având \* la final. Pentru evitarea overfitting-ului se folosește `cp(` parametru de complexitate), pentru graficul al doilea.



Putem observa din graficul de mai sus că cea mai mică valoare SSE (deviere) este atunci când  $\alpha=0,1$  și avem 9 noduri terminale.

Am creat apoi al doilea arbore cu  $cp=0$ , ceea ce înseamnă că toate împărțirile din setul de date sunt fără prunare, rezultând un arbore foarte complex pentru a vizualiza toate împărțirile posibile. Acesta prezintă cum în vreme ce complexitatea arborelui crește, eroarea relativă scade.



Locul unde se află cea mai mică eroare relativă este prezentat prin linia verticală punctată.

În continuare se caută cele mai bune valori pentru maxdepth si minsplit, de care depinde complexitatea arborelui nostru. Am folosit un hyper grid pentru a găsi cei mai buni parametri, folosind maxdepth între 8 și 15 și minsplit între 5 și 20.

```
> head(hyper_grid)
  minsplit maxdepth
1         5         8
2         6         8
3         7         8
4         8         8
5         9         8
6        10         8
```

Din imaginea prezentată vedem următoarele:

- dacă arborele are minim 5 observații pentru fiecare împărțire, va avea o adâncime de 8 și așa mai departe.

```
  minsplit maxdepth   cp   error
1        14        15 0.01 0.4910224
2         9        11 0.01 0.4958810
3        18        10 0.01 0.4963446
4        11        14 0.01 0.4967964
5        20        10 0.01 0.4971363
```

Se observă că arborele optim trebuie să aibă cel puțin 14 instanțe și o adâncime de 15 niveluri. Pentru prevenirea overfittingului, trebuie să avem un cp de 0,01. Se vede și că în acest caz, cea mai mică eroare relativă (cross-validare) este de 0,4910224.

De asemenea, am obținut valoarea RMSE-ului de 254806,5, indicând faptul că modelul are o precizie medie în ceea ce constă prețul locuințelor în funcție de suprafața locuinței, numărul de băi, numărul de camere, numărul de etaje și anul construcției. Așadar, modelul e unul relativ bun, dar trebuie să ținem cont că predicțiile nu sunt întotdeauna reale.

## Concluzie

Acest proiect s-a concentrat asupra întrebării din introducere și anume, „Cum influențează numărul dormitoarelor, suprafața în metri pătrați (  $m^2$ ), numărul băilor, etajele și anul construcției imobilului, prețul de cumpărare? “.

Am realizat cercetări pe baza variabilelor sqft\_living (suprafața locuibilă), bedrooms (numărul de dormitoare), bathrooms (numărul de băi), floors (numărul de etaje) și yr\_built (anul construcției imobilului). Am folosit un modele de regresie simplă, un model de regresie multiplă și un model de arbore de decizie. În urma realizării modelului de regresie multiplă, s-a constatat că factorii ce influențează cel mai mult prețul unei locuințe sunt suprafața locuibilă, numărul de băi și numărul de etaje. Având un R-squared de 55,51% înseamnă că modelul explică prețul locuințelor cu variabilele alese în proporție de 55,51%. Acesta are și un RMSE de 247177,7 aceasta fiind o diferență între prețurile prezise de model și cele reale. În cadrul arborelui de decizie, putem observa că cea mai importantă variabilă e, în mod sigur, suprafața locuinței, iar RMSE-ul aflat e de 254806,5. Acesta indică o valoare mai puțin mai mare decât cea de la modelul de regresie, indicând un model mai slab.

În concluzie, modelul de regresie multiplă este mai eficient și are performanțe mai bune.