IntroducereR_Proiect

Matei

2026-01-07

Exploratory Analysis

Introducere

Obiectivul proiectului:

- Analiza mecanismelor de formare a pretului pe piata imobiliara din Iasi. Cercetarea se concentreaza pe identificarea si cuantificarea relatiilor de dependenta dintre pretul de vanzare al apartamentelor (variabila dependenta) si principalii factori de influenta, cum ar fi suprafata utila, numarul de camere (variabile independente).

Vom folosi setul de date obtinut prin scraping si curatat in notebooks/2_data_cleaning_etl.ipynb (toti pasii urmati pentru a curata setul de date sunt explicati in acest notebook)

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

Incarcarea setului de date:

```
df <- read.csv('../data/processed/2_clean_data.csv')

head(df, n = 5)
```

```
##    X   Pret Suprafata_Utila Camere Pret_mp           Zona   Vechime_Imobil
## 1  0  72500            41.0      1 1768.29   Nicolina-CUG  Nou (Post-2000)
## 2  2 180000            72.4      3 2486.19   Centru-Civic  Nou (Post-2000)
## 3 11 106000            50.0      2 2120.00   Nicolina-CUG  Nou (Post-2000)
## 4 12  73500            57.0      2 1289.47 Tatarasi-Tudor Vechi (Pre-1977)
## 5 17 105000            80.0      3 1312.50         Bucium  Nou (Post-2000)
##   Compartimentare              Tip_Zona
## 1      Decomandat Accesibil/Rezidential
## 2      Decomandat               Premium
## 3      Decomandat Accesibil/Rezidential
## 4  Semidecomandat        Standard/Urban
## 5    Nespecificat             Periferie
```

Prezentarea bazei de date

Datele folosite sunt reale, obtinute din anunturi imobiliare de pe Storia, prin web scraping.

Esantionul cu care vom lucra contine 520 de observatii si 7 variabile.

Structura bazei de date:

- **Pret (numeric)**: pretul apartamentului, in euro. Este variabila pe care dorim sa o explicam.
- **Suprafata_Utila (numeric)**: suprafata apartamentului, in metri patrati. In mod intuitiv, se asteapta o corelatie pozitiva puternica cu pretul.
- **Camere (numeric/factor)**: numarul de camere. Variabila va fi tratata ca numeric pentru a permite calculul coeficientilor de corelatie bivariata si testarea liniaritatii.
- **Pret_mp (numeric)**: pretul pe metru patrat, obtinut prin (Pret / Suprafata Utila)
- **Zona (factor)**: o zona din Iasi, de exemplu "Nicolina-CUG", "Bucium", "Alexandru-Dacia"
- **Vechime_Imobil (factor)**: variabila obtina prin transformarea unei variabile numerice intr-o variabila categoriala, cu cel putin 3 categorii: "Vechi (Pre-1977)", "Clasic (1977-2000)", "Nou (Post-2000)"
- **Compartimentare (factor)**: Decomandat, Semidecomandat, Nespecificat. Vom elimina aceasta variabila imediat.
- **Tip_Zona (factor)**: Premium ('Copou-Saras', 'Centru-Civic'), Standard/Urban ('Tatarasi-Tudor', 'Podu-Ros-Cantemir', 'Pacurari-Canta', 'Alexandru-Dacia'), Accesibil/Rezidential ('Nicolina-CUG', 'Galata-Frumoasa')

In cazul de fata coloana **X** contine indicii ramasi in urma curatarii deci va trebui sa o scoatem. Coloana **Compartimentare** a fost pastrata pentru a demonstra limitarile scraping-ului, foarte multe observatii fiind de tipul "Nespecificat" deoarece scraping-ul s-a realizat pe pagina principala cu anunturile si descrierile anunturilor erau doar un preview. Ca sa fi accesat intreaga descriere trebuie deschis fiecare anunt in parte si extras continutul.

```
names(df)
```

```
## [1] "X"               "Pret"            "Suprafata_Utila" "Camere"
## [5] "Pret_mp"         "Zona"            "Vechime_Imobil"  "Compartimentare"
## [9] "Tip_Zona"
```

```
glimpse(df)
```

```
## Rows: 520
## Columns: 9
## $ X               <int> 0, 2, 11, 12, 17, 21, 28, 31, 36, 48, 51, 54, 55, 56, ~
## $ Pret            <int> 72500, 180000, 106000, 73500, 105000, 72000, 99000, 12~
## $ Suprafata_Utila <dbl> 41.00, 72.40, 50.00, 57.00, 80.00, 40.00, 54.00, 53.00~
## $ Camere          <int> 1, 3, 2, 2, 3, 1, 3, 2, 3, 3, 3, 2, 2, 3, 2, 2, 1, 2, ~
## $ Pret_mp         <dbl> 1768.29, 2486.19, 2120.00, 1289.47, 1312.50, 1800.00, ~
## $ Zona            <chr> "Nicolina-CUG", "Centru-Civic", "Nicolina-CUG", "Tatar~
## $ Vechime_Imobil  <chr> "Nou (Post-2000)", "Nou (Post-2000)", "Nou (Post-2000)~
## $ Compartimentare <chr> "Decomandat", "Decomandat", "Decomandat", "Semidecoman~
## $ Tip_Zona        <chr> "Accesibil/Rezidential", "Premium", "Accesibil/Reziden~
```

```
dim(df)
```

```
## [1] 520   9
```

Eliminam coloanele mentionate mai sus:

```
drops <- c("X", "Compartimentare")
df <- df[,!(names(df) %in% drops )]
```

```
dim(df)
```

```
## [1] 520   7
```

```
glimpse(df)
```

```
## Rows: 520
## Columns: 7
## $ Pret            <int> 72500, 180000, 106000, 73500, 105000, 72000, 99000, 12~
## $ Suprafata_Utila <dbl> 41.00, 72.40, 50.00, 57.00, 80.00, 40.00, 54.00, 53.00~
## $ Camere          <int> 1, 3, 2, 2, 3, 1, 3, 2, 3, 3, 3, 2, 2, 3, 2, 2, 1, 2, ~
## $ Pret_mp         <dbl> 1768.29, 2486.19, 2120.00, 1289.47, 1312.50, 1800.00, ~
## $ Zona            <chr> "Nicolina-CUG", "Centru-Civic", "Nicolina-CUG", "Tatar~
## $ Vechime_Imobil  <chr> "Nou (Post-2000)", "Nou (Post-2000)", "Nou (Post-2000)~
## $ Tip_Zona        <chr> "Accesibil/Rezidential", "Premium", "Accesibil/Reziden~
```

Vrem sa transformam variabilele de tip chr in factor, pentru a le trata corect in analizele statistice si in modelele de regresie. “Zona”, “Vechime_Imobil”, “Tip_Zona” sunt gandite sa fie de tip factor.

```
df$Zona <- as.factor(df$Zona)
df$Vechime_Imobil <- as.factor(df$Vechime_Imobil)
df$Tip_Zona <- as.factor(df$Tip_Zona)
```

```
glimpse(df)
```

```
## Rows: 520
## Columns: 7
## $ Pret            <int> 72500, 180000, 106000, 73500, 105000, 72000, 99000, 12~
## $ Suprafata_Utila <dbl> 41.00, 72.40, 50.00, 57.00, 80.00, 40.00, 54.00, 53.00~
## $ Camere          <int> 1, 3, 2, 2, 3, 1, 3, 2, 3, 3, 3, 2, 2, 3, 2, 2, 1, 2, ~
## $ Pret_mp         <dbl> 1768.29, 2486.19, 2120.00, 1289.47, 1312.50, 1800.00, ~
## $ Zona            <fct> Nicolina-CUG, Centru-Civic, Nicolina-CUG, Tatarasi-Tud~
## $ Vechime_Imobil  <fct> Nou (Post-2000), Nou (Post-2000), Nou (Post-2000), Vec~
## $ Tip_Zona        <fct> Accesibil/Rezidential, Premium, Accesibil/Rezidential,~
```

Mult mai bine, acum tipul variabilelor este cel corect.

```
summary(df$Pret)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   38000   78875   98700  100079  121625  194900
```

```
summary(df$Suprafata_Utila)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   20.00   43.00   53.00   53.73   64.00   91.61
```

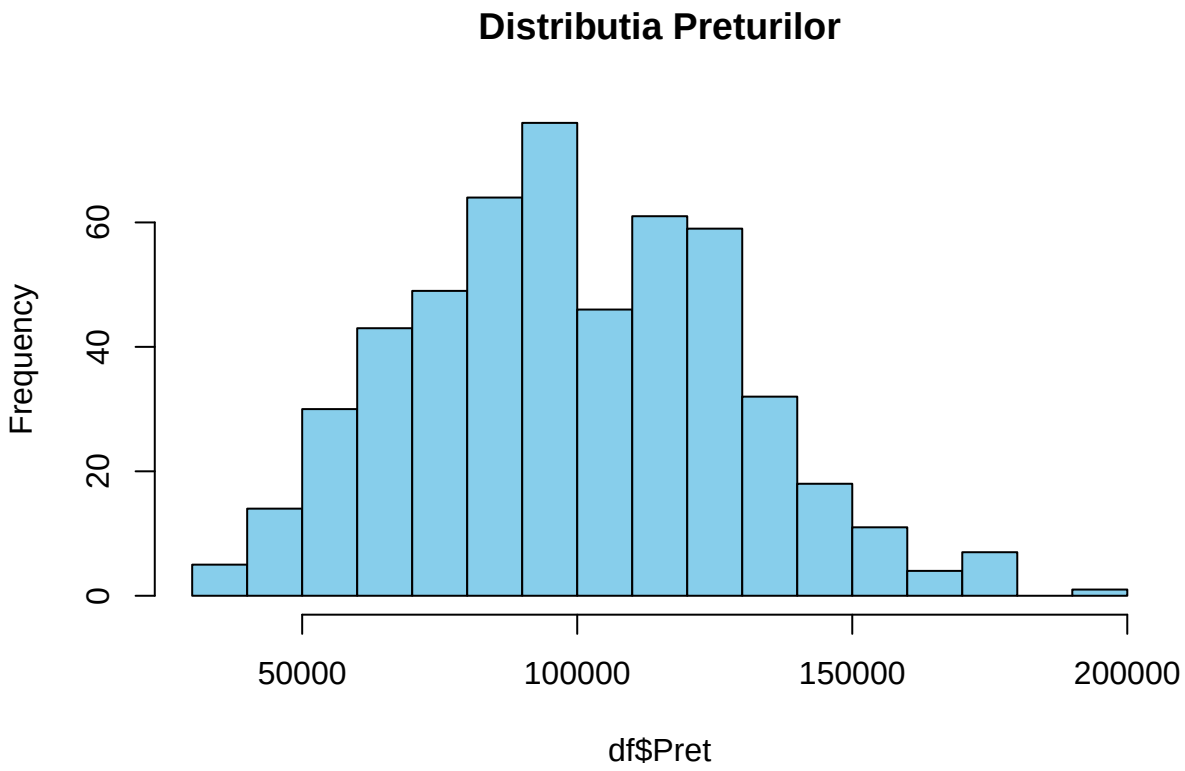
```
summary(df$Camere)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   2.000   2.000   2.133   3.000   3.000
```

OBS:

- Avem apartamente pana in 3 camere
- Setul de date a fost explorat si curatat anterior

```r
hist(df$Pret, breaks=20, main="Distributia Preturilor", col="skyblue")
```



Distributia este apropiata de una normala, dar usor asimetrica spre dreapta.

Traducere:

- Sunt cateva preturi mai mari care trag coada spre dreapta

Zona centrala (intervalul de pret standard/cel mai frecvent):

- Majoritatea valorilor sunt intre 70-130 de mii
- Varful distributiei pare in jur de 90-100 de mii

Valori extreme/outlieri:

- Exista cateva valori mici (40-50 mii), dar si valori foarte mari (160-200 mii)
- Nu par outlieri extrem de agresivi (nu sunt erori), dar preturile mari sunt mai rare

Implicatii statistice:

- Media probabil este putin mai mare decat mediana, din cauza asimetriei spre dreapta

```r
mean(df$Pret)
```

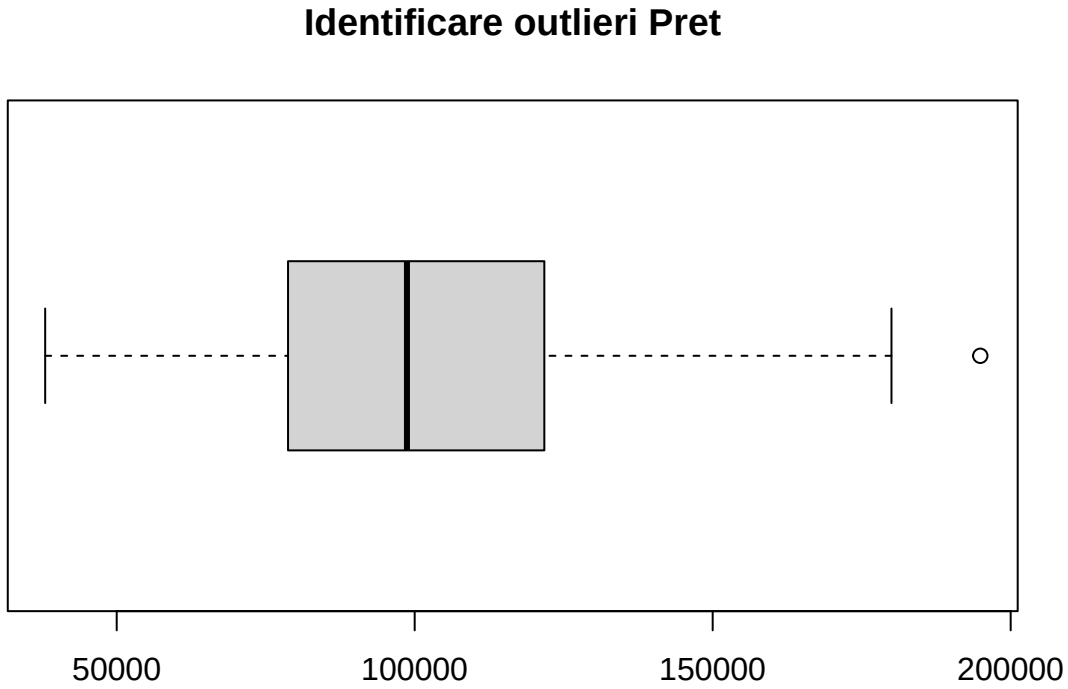
```
## [1] 100079.2
```

```r
median(df$Pret)
```

```
## [1] 98700
```

Se confirma ideea, media fiind 100 mii in timp ce mediana 98 de mii

```r
boxplot(df$Pret, main = "Identificare outlieri Pret", horizontal = T)
```



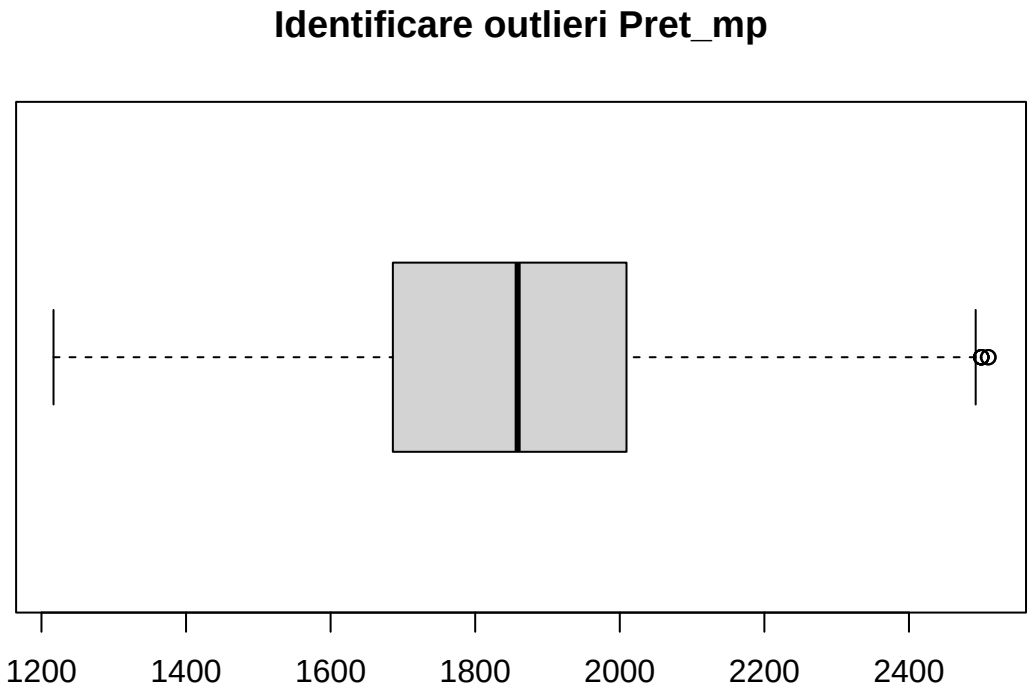
In curatarea anterioara s-a folosit Interquartile Range (IQR) pentru detectarea outlierilor de Pret.

Aceeasi metoda s-a folosit si pentru detectarea outlierilor de Suprafata_Utila, Camere, Pret_mp.

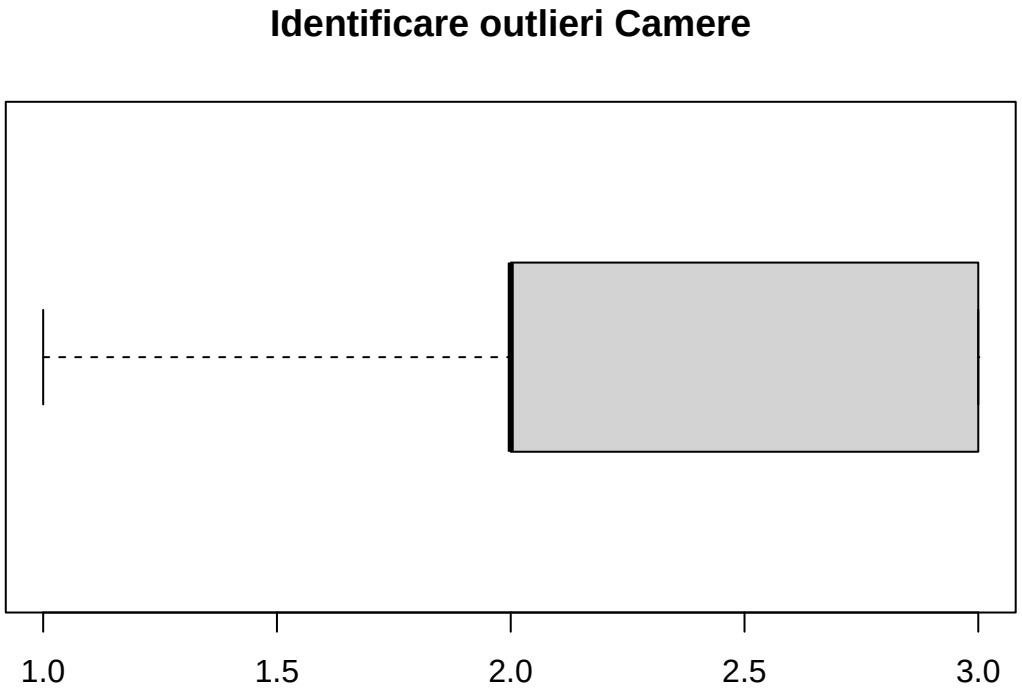
Au fost eliminate apartamentele cu o suprafata utila mai mica de 40mp, dar cu un pret mai mare de 75 de mii de euro (apartamente mici, dar neobisnuit de scumpe). La fel si apartamentele cu o suprafata utila mai mare de 80mp, dar cu un pret mai mic de 130 mii (apartamente mari, dar neobisnuit de ieftine).

Aceasta etapa documentata in fisierul *2_data_cleaning_etl.ipynb* (sectiunile IQR si Outlieri) a avut ca scop cresterea robustetii modelelor de regresie prin eliminarea zgomotului din date.

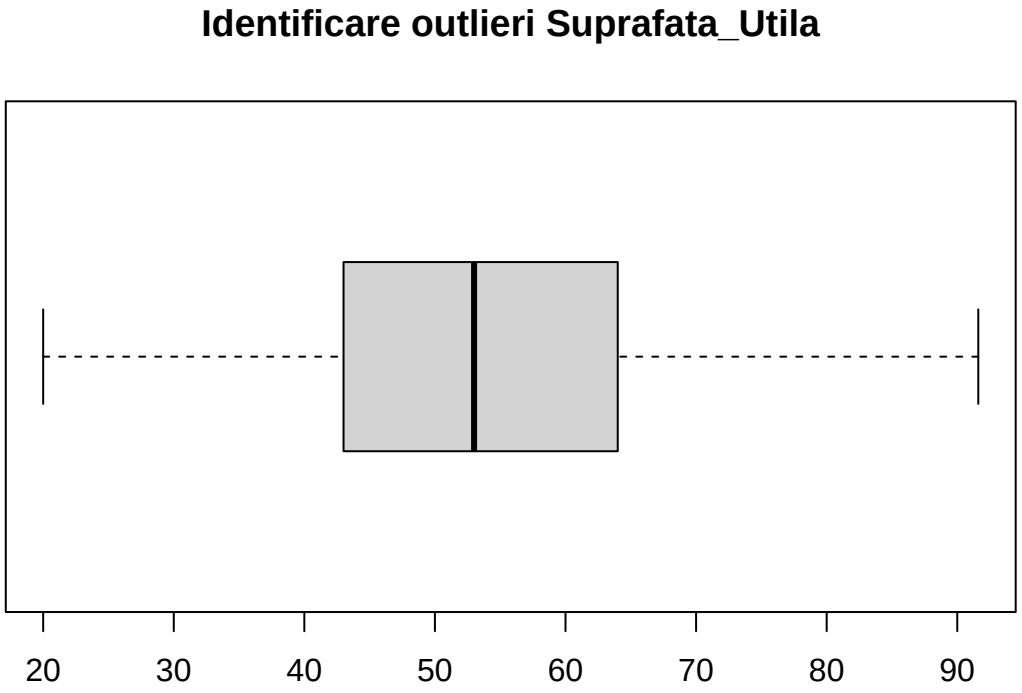
```
boxplot(df$Pret_mp, main = "Identificare outlieri Pret_mp", horizontal = T)
```



```
boxplot(df$Camere, main = "Identificare outlieri Camere", horizontal = T)
```

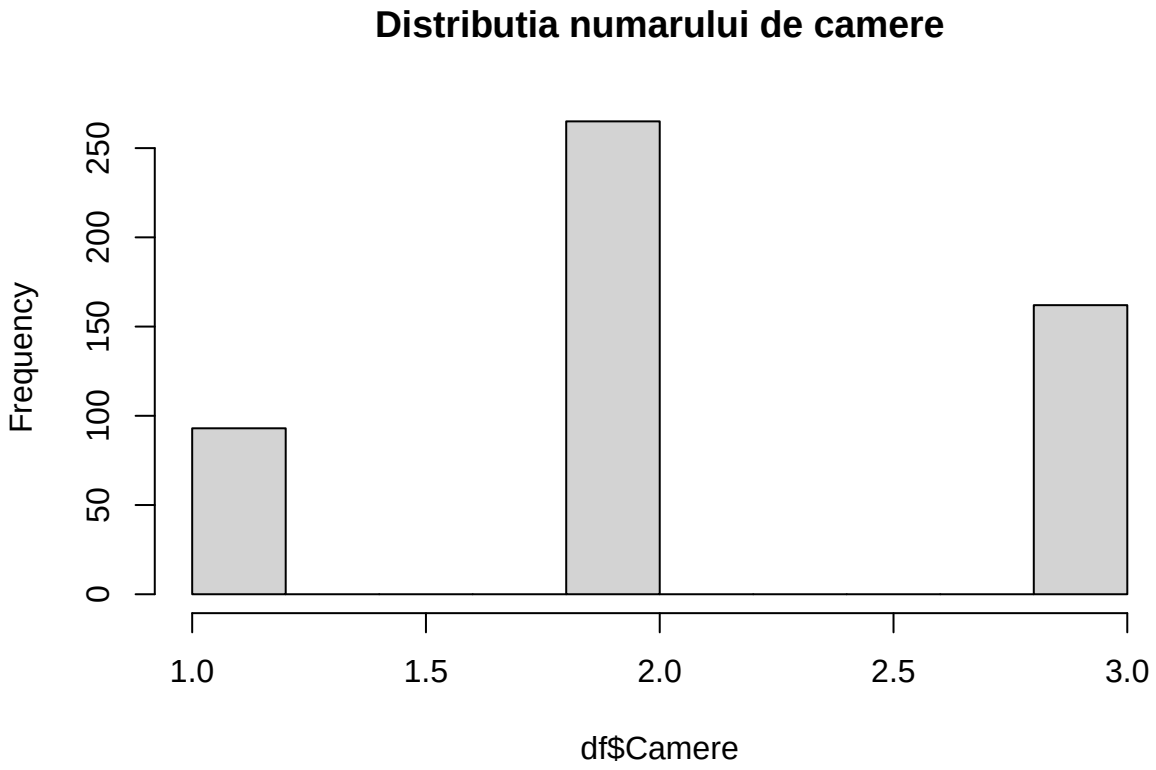



```
boxplot(df$Suprafata_Utila, main = "Identificare outlieri Suprafata_Utila", horizontal = T)
```



hist(df$Pret, breaks=20, main="Distributia Preturilor", col="skyblue")

```
hist(df$Camere, breaks = 10, main="Distributia numarului de camere")
```



Piata este dominata de apartamente cu 2 camere, cele cu 3 camere fiind a doua cea mai frecventa categorie.

Interpretare practica:

- Oferta in Iasi este orientata catre:
  - Cupluri/familii mici (2 camere)
  - Intr-o masura mai mica, familii mai mari (3 camere)

Frecventa absoluta:

```
table(df$Camere)
```

```
##
##   1   2   3
##  93 265 162
```

Frecventa relativa:

```
prop.table(table(df$Camere))
```

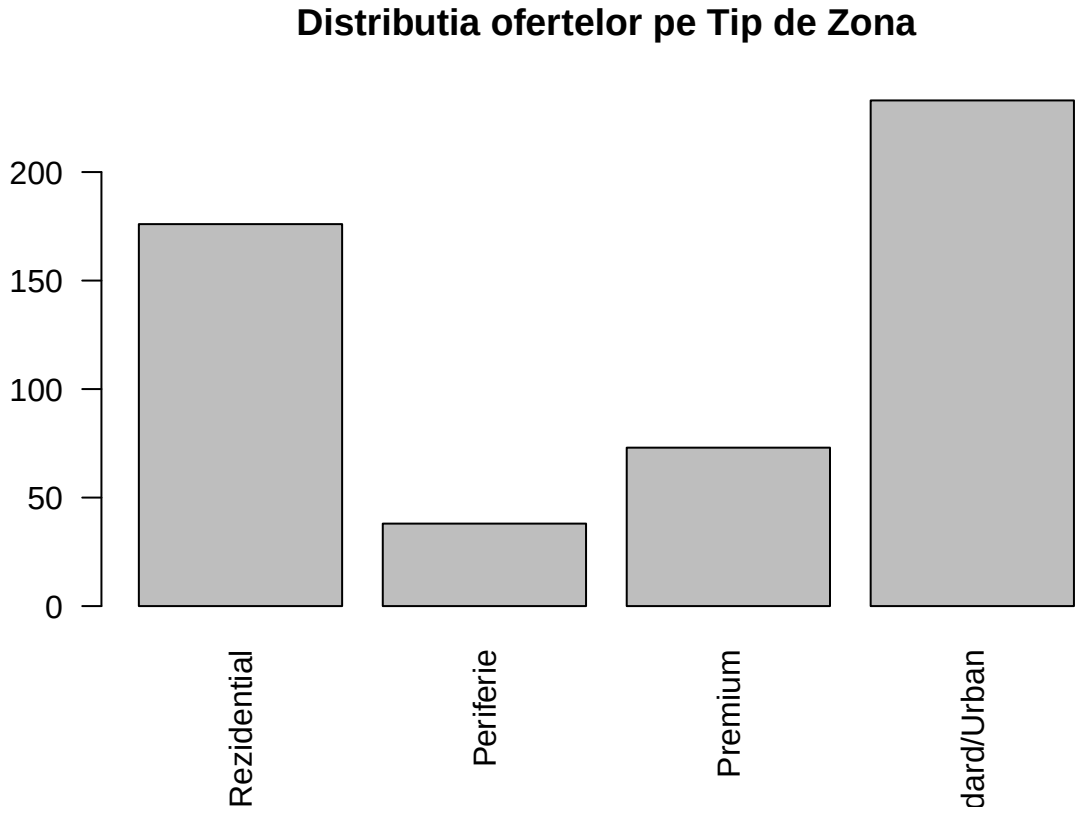
```
##
##         1         2         3
## 0.1788462 0.5096154 0.3115385
```

Jumatate (50%) din numarul de oferte sunt apartamente de 2 camere.

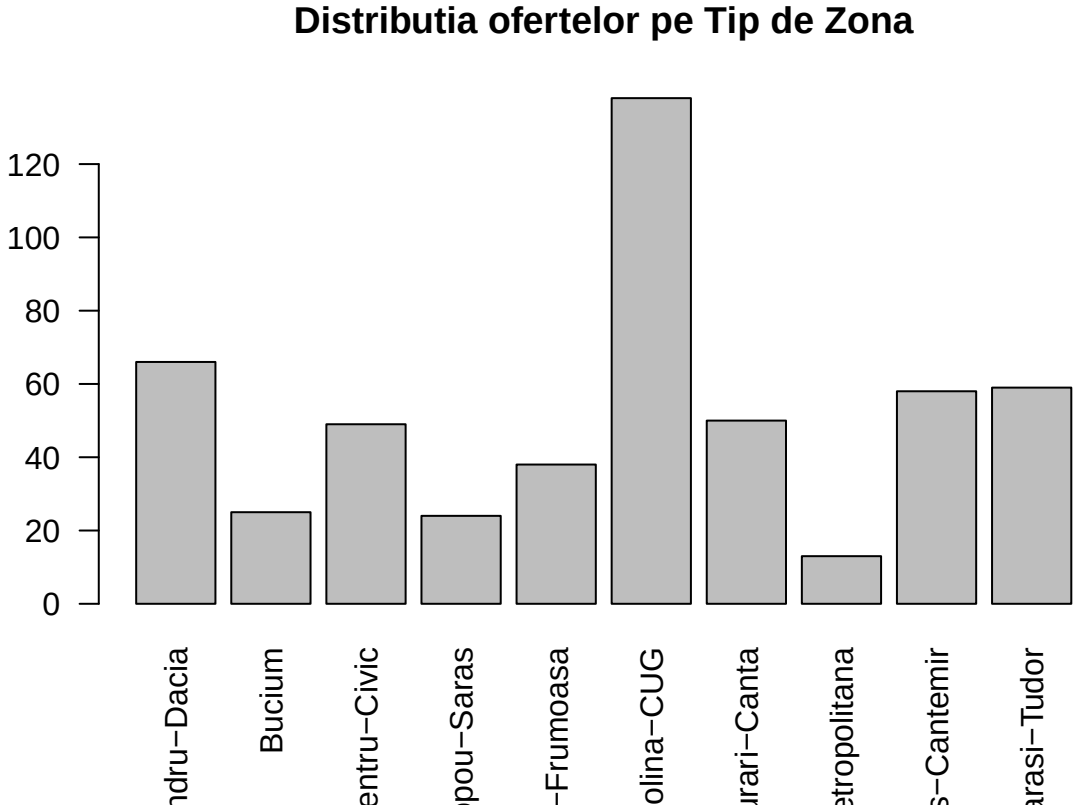
Analiza variabilelor categoriale

Analiza frecventei pentru Tip_Zona:

```
tab_zona <- table(df$Tip_Zona)
barplot(tab_zona, main="Distributia ofertelor pe Tip de Zona", las=2)
```




```
tab_zona <- table(df$Zona)
barplot(tab_zona, main="Distributia ofertelor pe Tip de Zona", las=2)
```



Se observa cum pentru unele zone nu sunt destule observatii, cum ar fi “Periferie-Metropolitana”, “Bucium”.

Cum s-a realizat maparea la zone:

‘Nicolina-CUG’: [‘Nicolina 1’, ‘Nicolina 2’, ‘CUG’, ‘Hlincea’, ‘Tudor Neculai’, ‘Soseaua Nicolina’, ‘Poitiers’, ‘Manta Rosie’],

‘Centru-Civic’: [‘Centru’, ‘Palas’, ‘Independentei’, ‘Academiei’, ‘Ion Creanga’, ‘Carol I’, ‘Anastasie Panu’, ‘Cuza Voda’, ‘Arcu’, ‘Smardan’, ‘Podu de Fier’],

‘Podu-Ros-Cantemir’: [‘Podu Ros’, ‘Cantemir’, ‘Tesatura’],

‘Tatarasi-Tudor’: [‘Tatarasi Sud’, ‘Tatarasi Nord’, ‘Vasile Lupu’, ‘Oancea’, ‘Tudor Vladimirescu’, ‘Baza 3’],

‘Pacurari-Canta’: [‘Pacurari’, ‘Canta’, ‘Moara de Foc’],

‘Copou-Saras’: [‘Copou’, ‘Agronomie’, ‘Sadoveanu’, ‘Agronomilor’, ‘Moara de Vant’, ‘Ticau’],

‘Alexandru-Dacia’: [‘Alexandru Cel Bun’, ‘Dacia’, ‘Mircea cel Batran’, ‘Bularga’, ‘Decebal’],

‘Bucium’: [‘Bucium’, ‘Visan’, ‘Barnova’],

‘Galata-Frumoasa’: [‘Galata’, ‘Frumoasa’, ‘Ciurea’, ‘Bisericii’],

‘Periferie-Metropolitana’: [‘Miroslava’, ‘Rediu’, ‘Dancu’, ‘Aroneanu’, ‘Valea Lupului’, ‘Voinesti’]

Variabila “Tip_Zona”

- Premium: Copou-Saras, Centru-Civic
- Standard/Urban: Tatarasi-Tudor, Podu-Ros-Cantemir, Pacurari-Canta, Alexandru-Dacia
- Accesibil/Rezidential: Nicolina-CUG, Galata-Frumoasa
- Periferie: Bucium, Periferie-Metropolitana

De ce ?

In statistica, pentru ca un grup sa fie relevant, este recomandat sa aiba macar 30 de observatii sau macar un numar apropiat. In cazul nostru sunt cateva sub 30 de observatii, dar care ar putea fi grupate in zone mai generale pentru a creste numarul de observatii.

Proportii vechime imobil:

```
prop.table(table(df$Vechime_Imobil))
```

```
##
## Clasic (1977-2000)    Nou (Post-2000)   Vechi (Pre-1977)
##          0.4538462          0.3615385          0.1846154
```

Cea mai mare majoritate a blocurilor sunt clasice (perioada 1977-2000), urmand cele Noi (>2000)

Analiza de regresie

Regresie liniara simpla

Scopul analizei:

- Vrem sa vedem impactul suprafetei utile (variabila independenta) asupra pretului apartamentului (variabila dependenta)

```r
regresie_simpla <- lm(Pret ~ Suprafata_Utila, data = df)
```

```r
summary(regresie_simpla)
```

```
##
## Call:
## lm(formula = Pret ~ Suprafata_Utila, data = df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -46843  -9781   -326   8438  47396
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      6486.05    2523.17   2.571   0.0104 *
## Suprafata_Utila  1741.96      45.32  38.434   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15060 on 518 degrees of freedom
## Multiple R-squared:  0.7404, Adjusted R-squared:  0.7399
## F-statistic:  1477 on 1 and 518 DF,  p-value: < 2.2e-16
```

Interpretarea rezultatelor

1. **Estimarea si testarea raportului de determinatie (R squared):**
   - R squared = 0.7404, adica 74% din variatia pretului este explicata de suprafata utila
   - Testarea acestuia este confirmata de F-statistic: 1477 cu un p-value $<$ 2.2e-16, adica raportul de determinatie este semnificativ statistic (modelul este util)

2. **Scrierea ecuatiei estimate:**
   - Pret = 6486.05 + 1741.96 * Suprafata_Utila

3. **Interpretarea parametrilor:**
   - Intercept = 6486.05 - valoarea teoretica a pretului pentru o suprafata de 0 mp
   - Suprafata_Utila = 1741.96 - la fiecare mp adaugat, pretul apartamentului creste in medie cu 1741.96 euro

4. **Testarea semnificatiei parametrilor:**
   - Pentru coeficientul suprafetei, valoarea t = 38.434 si p $<$ 2e-16 $<$ 0.05, indica faptul ca suprafata este un predictor semnificativ

5. **Testarea modelului:**

Testarea semnificatiei globale (testul F)

- Ipoteza nula H0: Toti coeficientii (in afara de intercept) sunt 0 (modelul nu e bun)
- Ipoteza alternativa H1: Cel putin un coeficient este diferit de 0
- Rezultat: p-value $<$ 2.2e-16 $<$ 0.05, respingem H0.
- Concluzie: Cu un risc asumat de 5%, cel putin un coeficient este diferit de 0, modelul fiind valid statistic

Testarea ipotezelor pe reziduuri (Teste formale)

a. Testul pentru normalitate (Shapiro-Wilk)

```
shapiro.test(residuals(regresie_simpla))
```

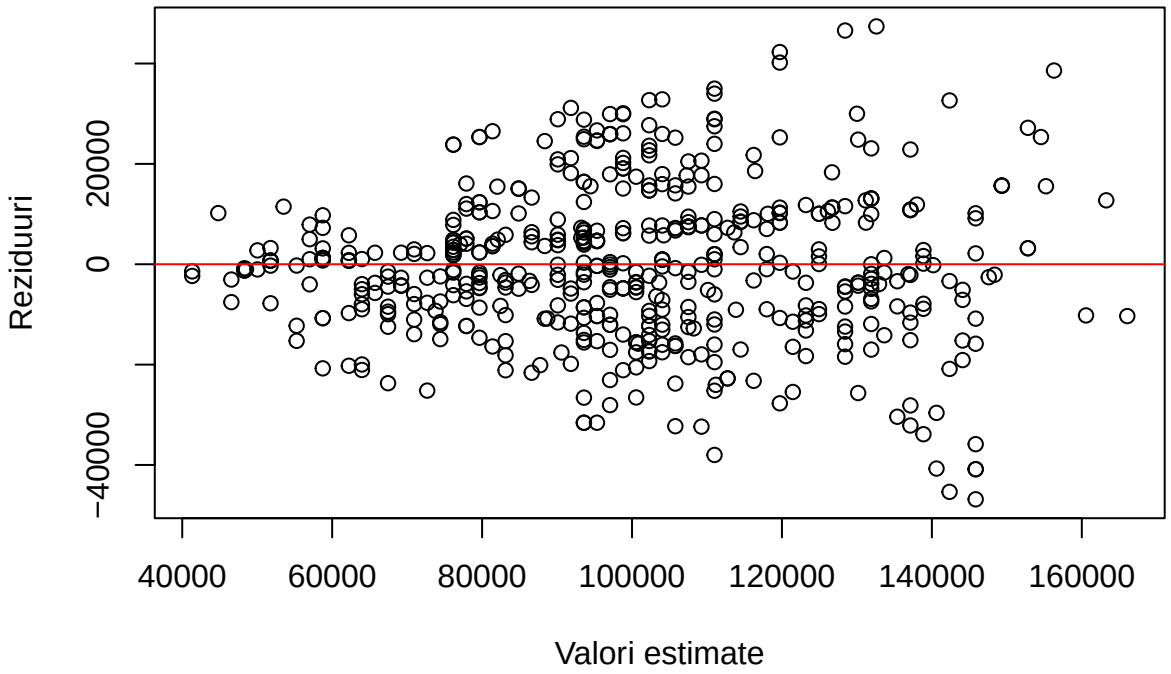
```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(regresie_simpla)
## W = 0.99365, p-value = 0.02801
```

- Ipoteza nula H0: reziduurile sunt distribuite normal
- Ipoteza alternativa H1: reziduurile nu sunt distribuite normal
- Rezultat: p-value = 0.02 < 0.05, respingem H0.
- Concluzie: Cu un risc asumat de 5%, reziduurile nu urmeaza o distributie normala, abaterea fiind relevanta din punct de vedere statistic
- W = 0.99365 apropiat de 1, ceea ce sugereaza abateri minore de la normalitate

Totusi, avand in vedere dimensiunea mare a esantionului (520 observatii) si faptul ca statistica W este apropiata de 1, abaterea de la normalitate este redusa si nu afecteaza semnificativ validitatea modelului.

b. Verificarea omoscedasticitatii (grafic reziduuri vs valori estimate)

```
plot(regresie_simpla$fitted.values, residuals(regresie_simpla),
     xlab = "Valori estimate",
     ylab = "Reziduuri")
abline(h = 0, col = "red")
```



Liniaritatea modelului:

- Reziduurile sunt distribuite in jurul valorii 0 (linia rosie)
- Nu apare un tipar clar (nu pare sa fie forma de U sau curba)
- Majoritatea punctelor sunt relativ simetric dispuse deasupra si sub axa 0

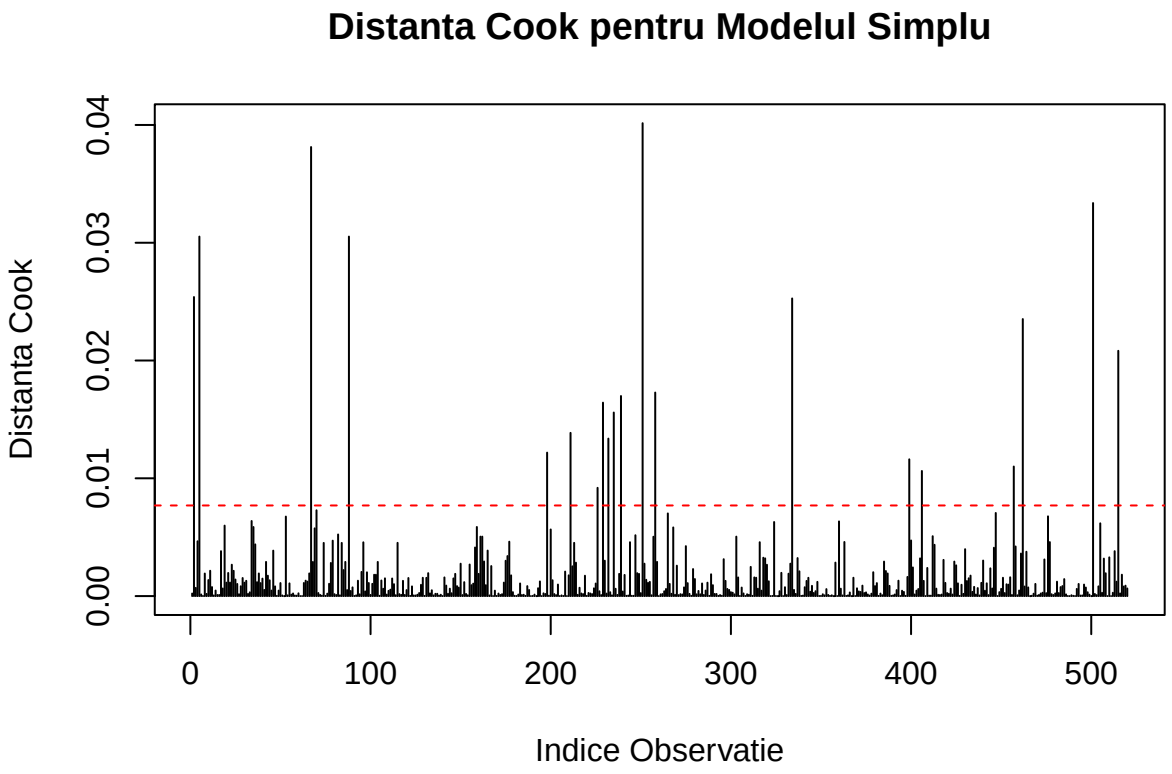
Problema minora:

- Pentru valori estimate mai mari (dreapta graficului), dispersia reziduurilor scade (usoara forma de evantai). Posibil ca variatia reziduurilor sa nu fie perfect constanta

6. Identificarea punctelor de influenta si grafice pentru reziduuri:

```
cooks_d <- cooks.distance(regresie_simpla)

plot(cooks_d, type = "h", main = "Distanta Cook pentru Modelul Simplu",
     ylab = "Distanta Cook", xlab = "Indice Observatie")
abline(h = 4/nrow(df), col = "red", lty = 2)
```



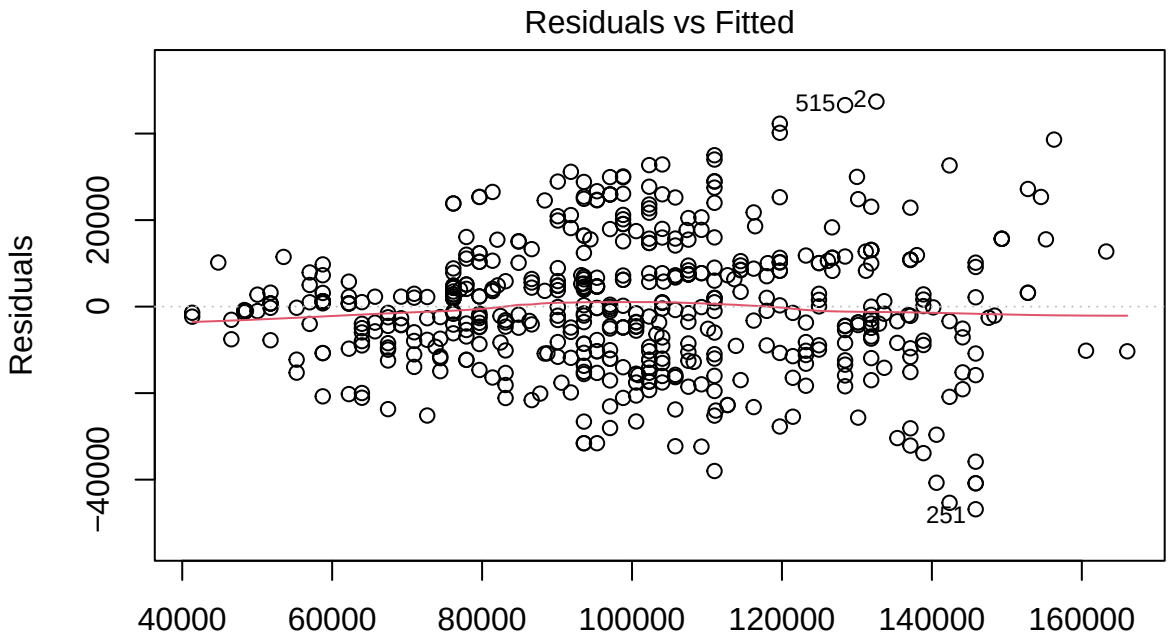
- Axa X (Indice Observatie) - fiecare apartament individual din baza mea de date (1-520)
- Axa Y (Distanta Cook) - masoara cat de mult se schimba coeficientii modelului daca acea observatie ar fi eliminata. Cu cat bara este mai inalta cu atat impactul acelui apartament asupra liniei de regresie este mai mare
- Linia rosie (Pragul Critic) - este setata la 4/n (n = 520). Orice bara care trece de aceasta linie este considerata un punct de influenta (din punct de vedere statistic).

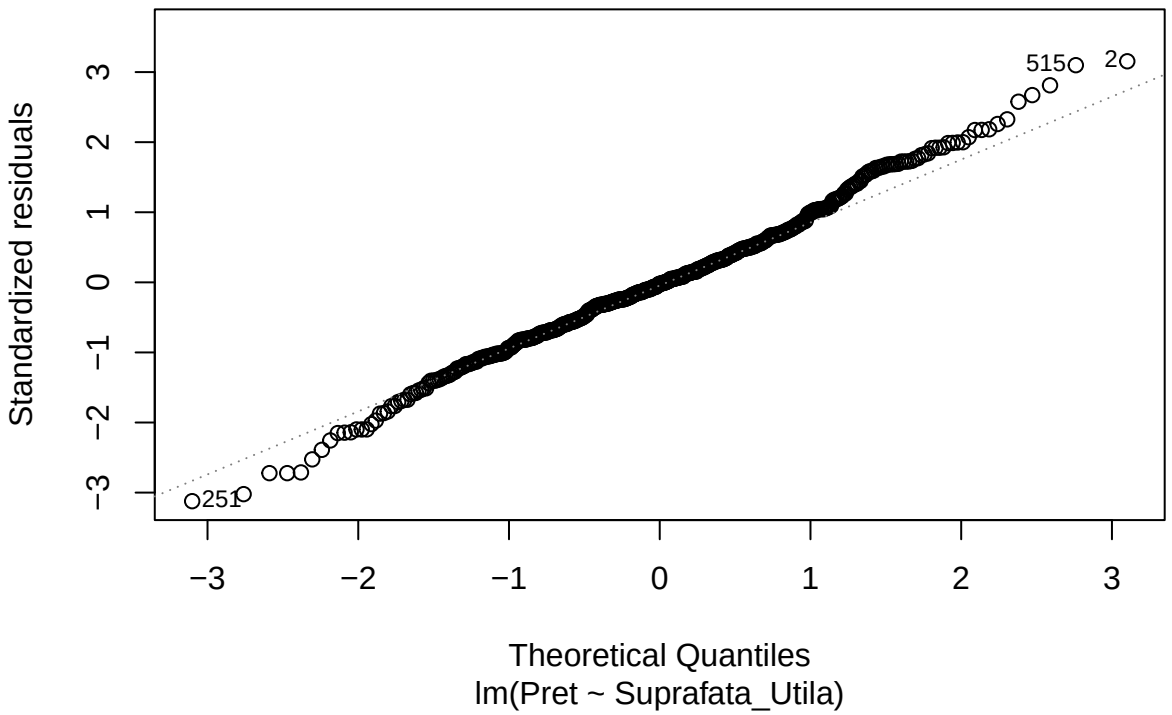
Undeva la 10-15 observatii sar mult peste linia rosie (pragul critic). Cu aproximatie 20 de observatii trec pragul.

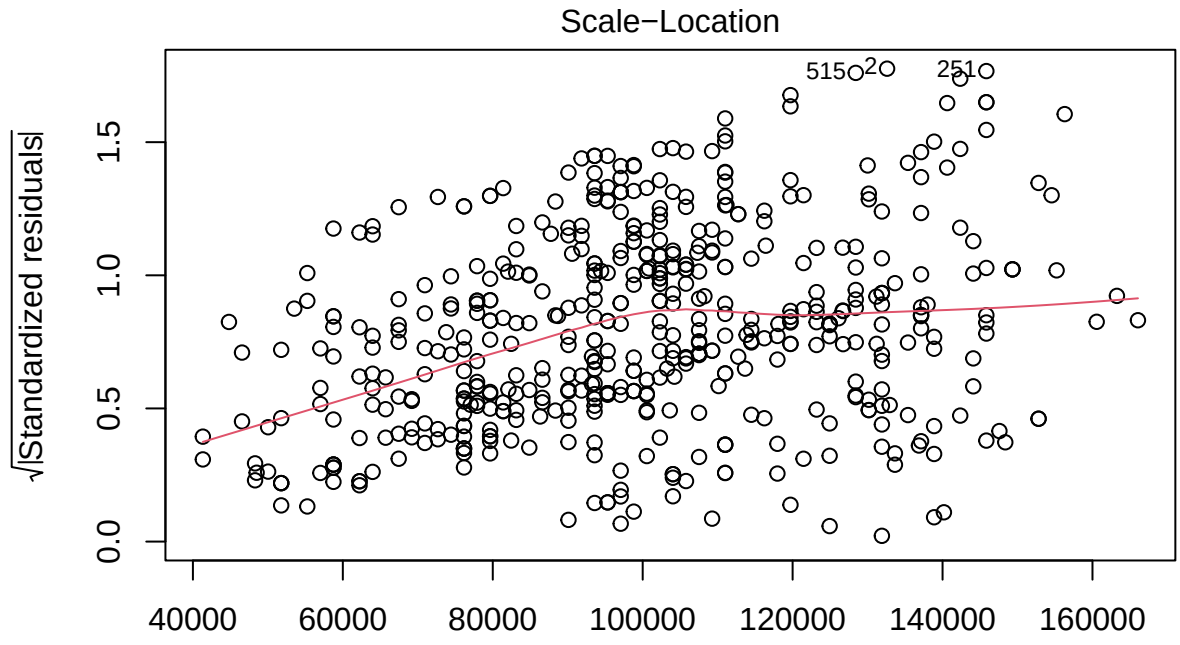
- Aceste puncte de influenta nu sunt neaparat erori, ci niste apartamente atipice. De exemplu:
  - Putem avea un apartament foarte mic dar intr-o zona extrem de buna, care are un pret de vanzare cu mult peste cat ar valora altele de aceeasi suprafata
  - Putem avea un penthouse de lux care are un pret atat de mare incat trage linia de regresie in sus, alterand estimarea pentru restul apartamentelor obisnuite

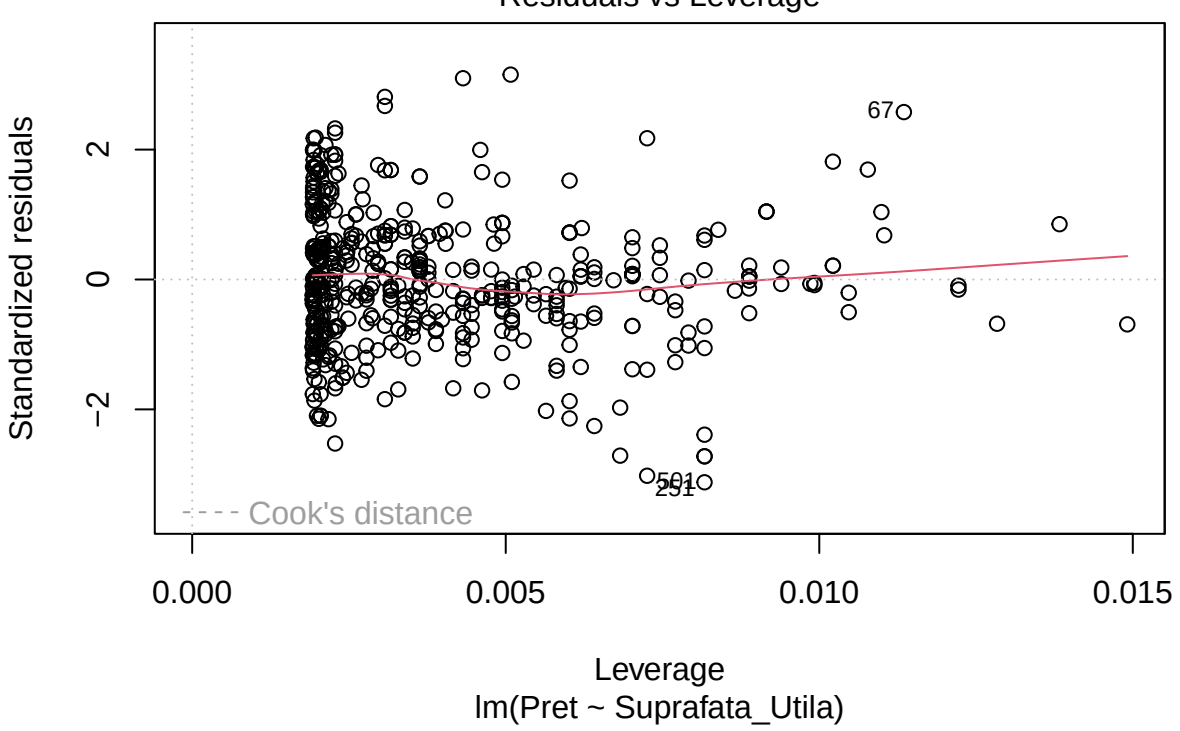
Vom pastra aceste valori deoarece reprezinta tranzactii reale pe piata din Iasi, insa prezenta lor explica de ce modelul ar putea avea erori mai mari in segmentele de pret extrem.

```
plot(regresie_simpla)
```









Important

Desi in primul grafic apareau multe observatii peste linia rosie, graficul Residuals vs Leverage demonstreaza ca acele puncte nu strica modelul:

- Absenta zonelor critice: Toate observatiile sunt grupate in partea stanga, departe de zonele de pericol din colturile din dreapta sus/jos

- Punctele etichetate (67, 251, 501): R eticheteaza automat cele mai extreme puncte. Chiar daca observatia 67 are un leverage mai mare si observatiile 251 si 501 au reziduuri mai mari, ele nu au o combinatie destul de puternica de ambele pentru a devia linia de regresie
- Linia rosie: este aproape perfect orizontala si suprapusa peste linia punctata de la 0. Acest lucru indica faptul ca nu exista un tipar sistematic de eroare si ca modelul este stabil.

Asadar, modelul este robust si poate fi folosit pentru predictie fara sa eliminam observatiile care depasesc pragul teoretic de 4/n.