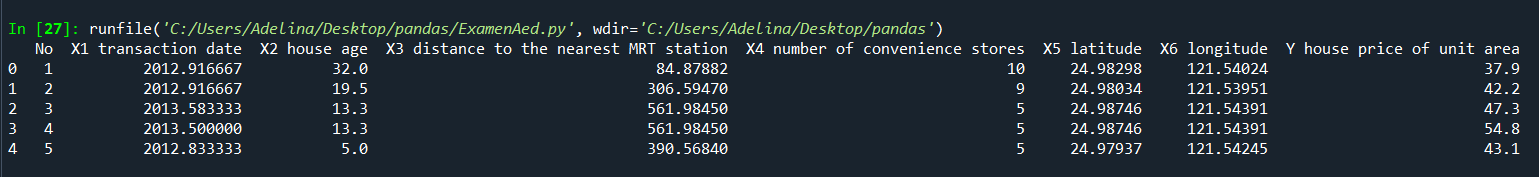
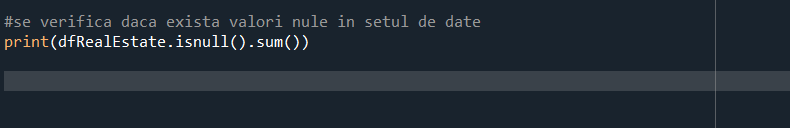
**Proiect: analiza si exploatarea datelor**

1. **Descrierea setului de date:**

* Setul de date Real Estate Valuation a fost colectat și donat către UCI de către Prof. I-Cheng Yeh, profesor în cadrul Universității Tamkang din Taiwan.
* Acest set de date conține date istorice despre piața imobiliară din districtul Sindian,orașul New Taipei,Taiwan, în perioada 2012-2013.
* Setul conține 414 înregistrări și 7 atribute,unde fiecare rând reprezintă o înregistrare a unei tranzacții de proprietate imobiliară, în timp ce fiecare coloană reprezintă un atribut care descrie acea înregistrare.



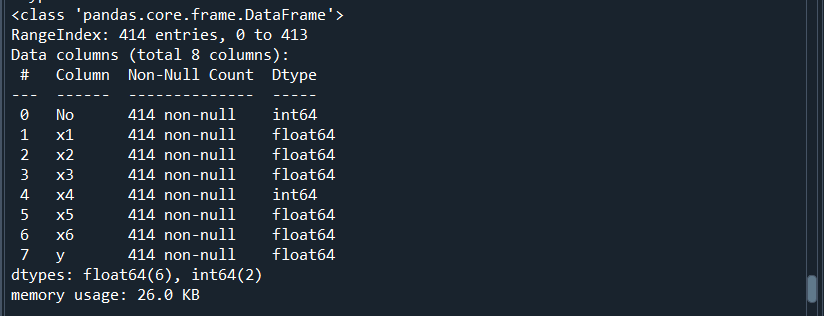
**2. Gestionarea datelor lipsă**





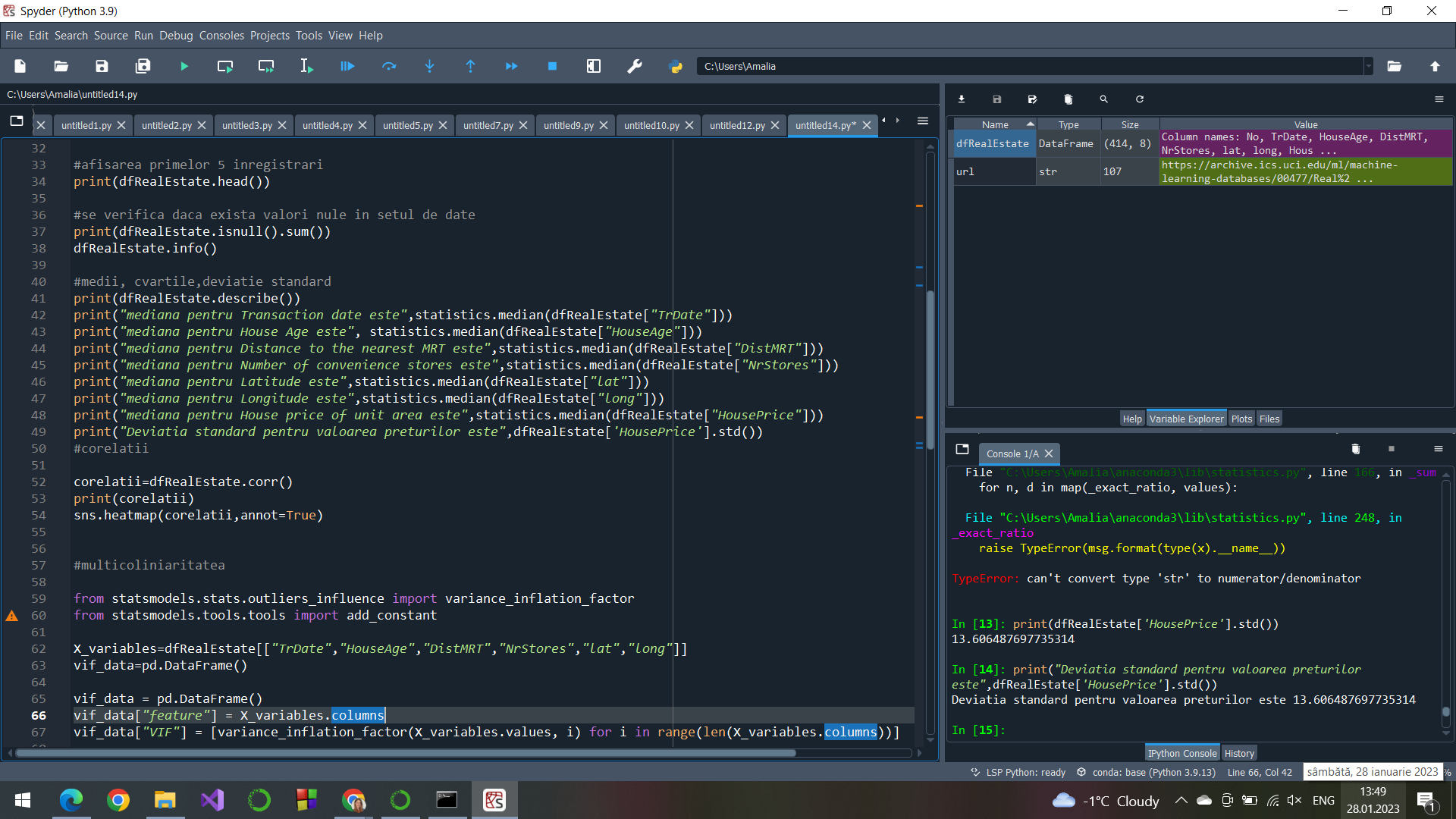
În acest set de date nu există date de tip null, adică toate câmpurile sunt completate cu valori, astfel nu e nevoie să folosi funcția **drop()** pentru a elimina aceste date lipsă.

Pentru a confirma rezultatul anterior, am folosit funcția **info()**. Astfel, s-a confirmat dublu că nu există valori nule în setul de date Real Estate Value. Observăm că tipul de date folosit pentru coloanele “No” și “X4 number of convenience stores” este int64, pentru restul float64.

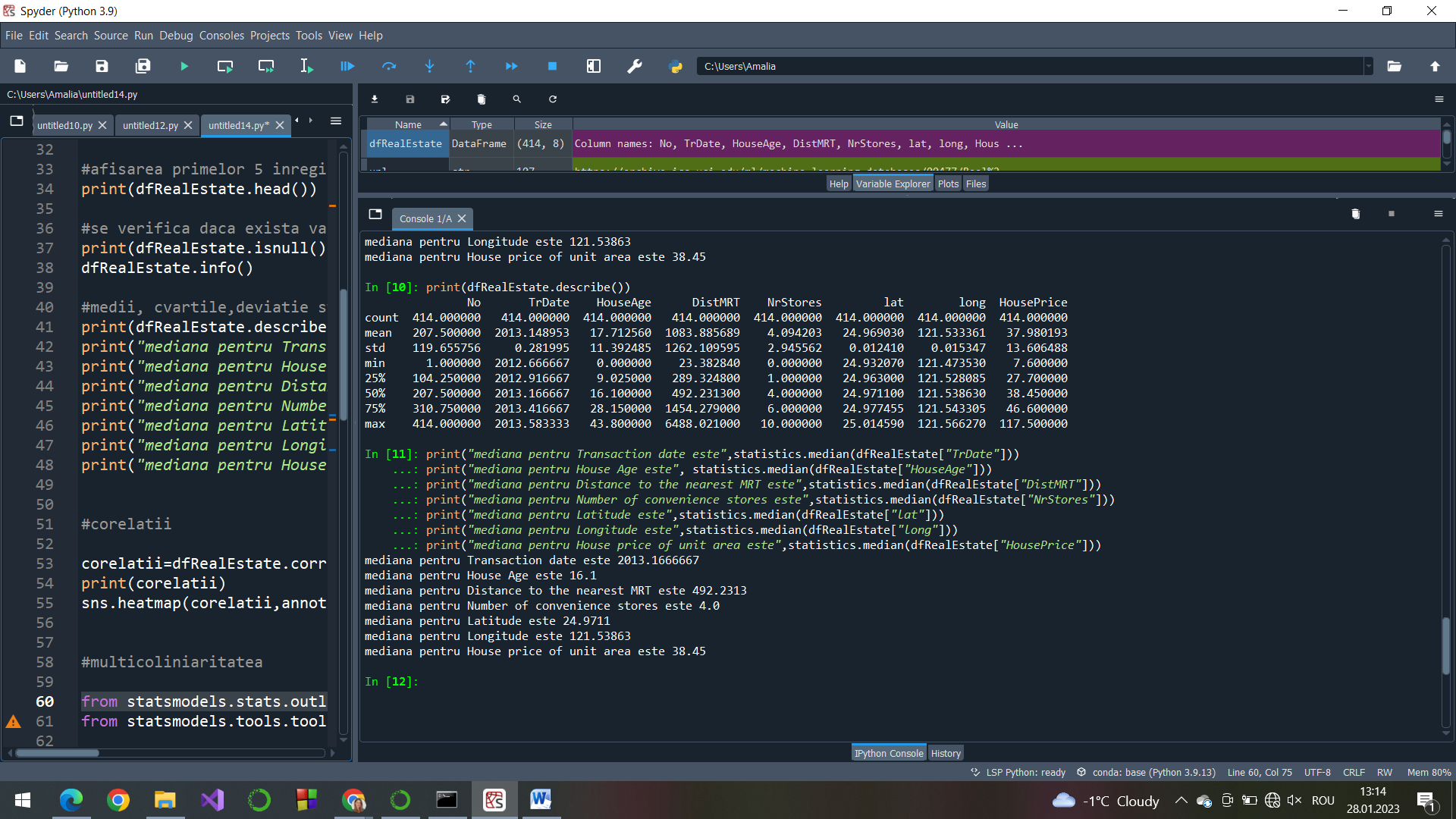


**3. Medii, cvartile, mediane, deviații standard**

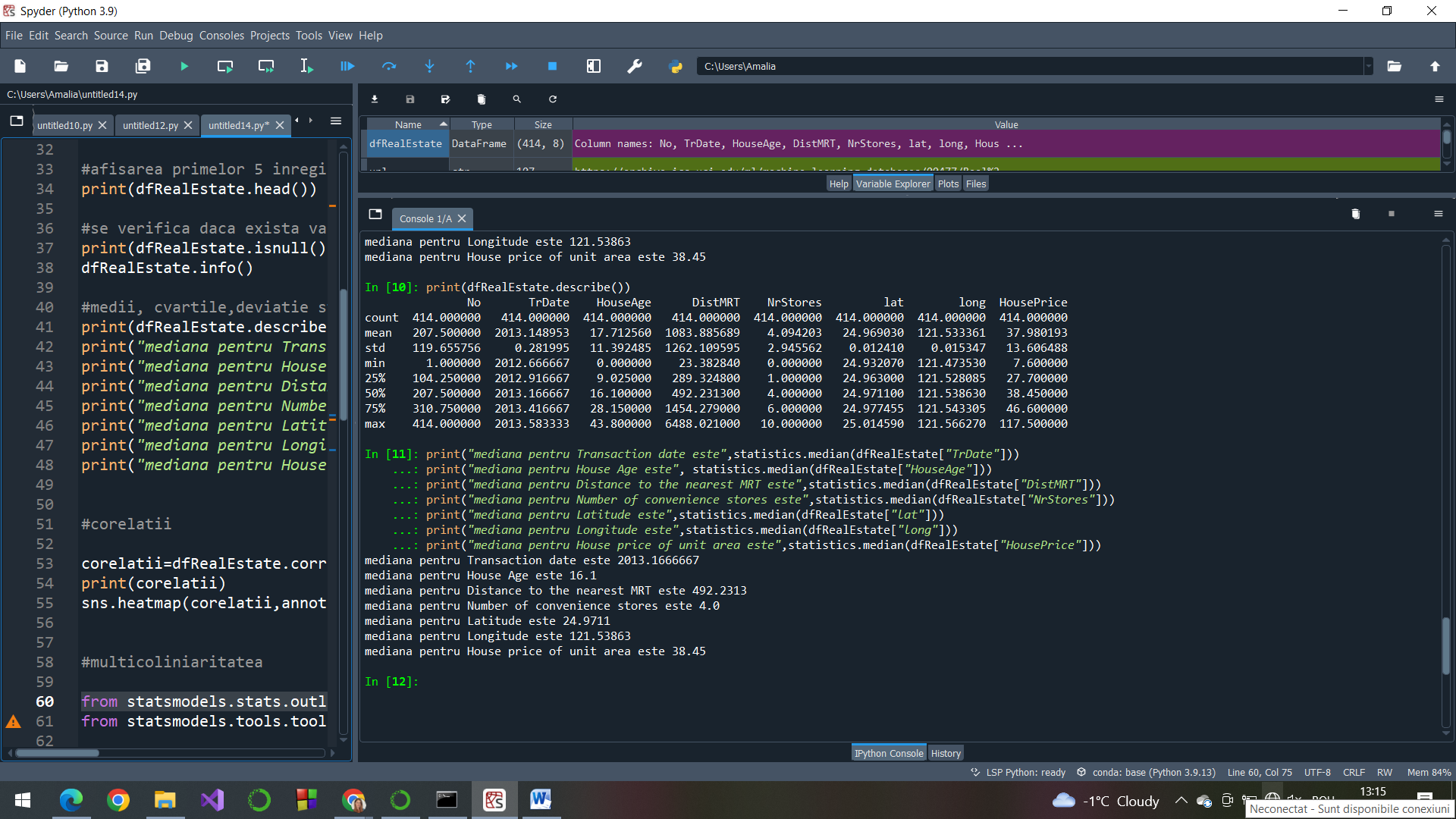
* **Media** se calculează prin adunarea tuturor valorilor din matricea de date, care este apoi împărțită la numărul de observații.
* **Cvartilă** măsoară răspândirea valorilor peste și sub medie prin împărțirea distribuției în patru grupuri. O cvartilă împarte datele în trei puncte – o *cvartilă inferioară*, mediană și *cvartilă superioară* – pentru a forma patru grupuri ale setului de date.
* **Mediana** este valoarea **medie** exactă a setului de date. Poate fi calculată prin aranjarea setului de date în ordine crescătoare și apoi găsirea sau alegerea valorii **medii** din setul de date.
* **Deviația standard** este folosită pentru a măsura dispersia sau răspândirea unui set de valori în jurul valorii medii. Astfel, cu cât abaterea standard este mai mică, cu atât este mai omogenă “populația” respectivului lot.

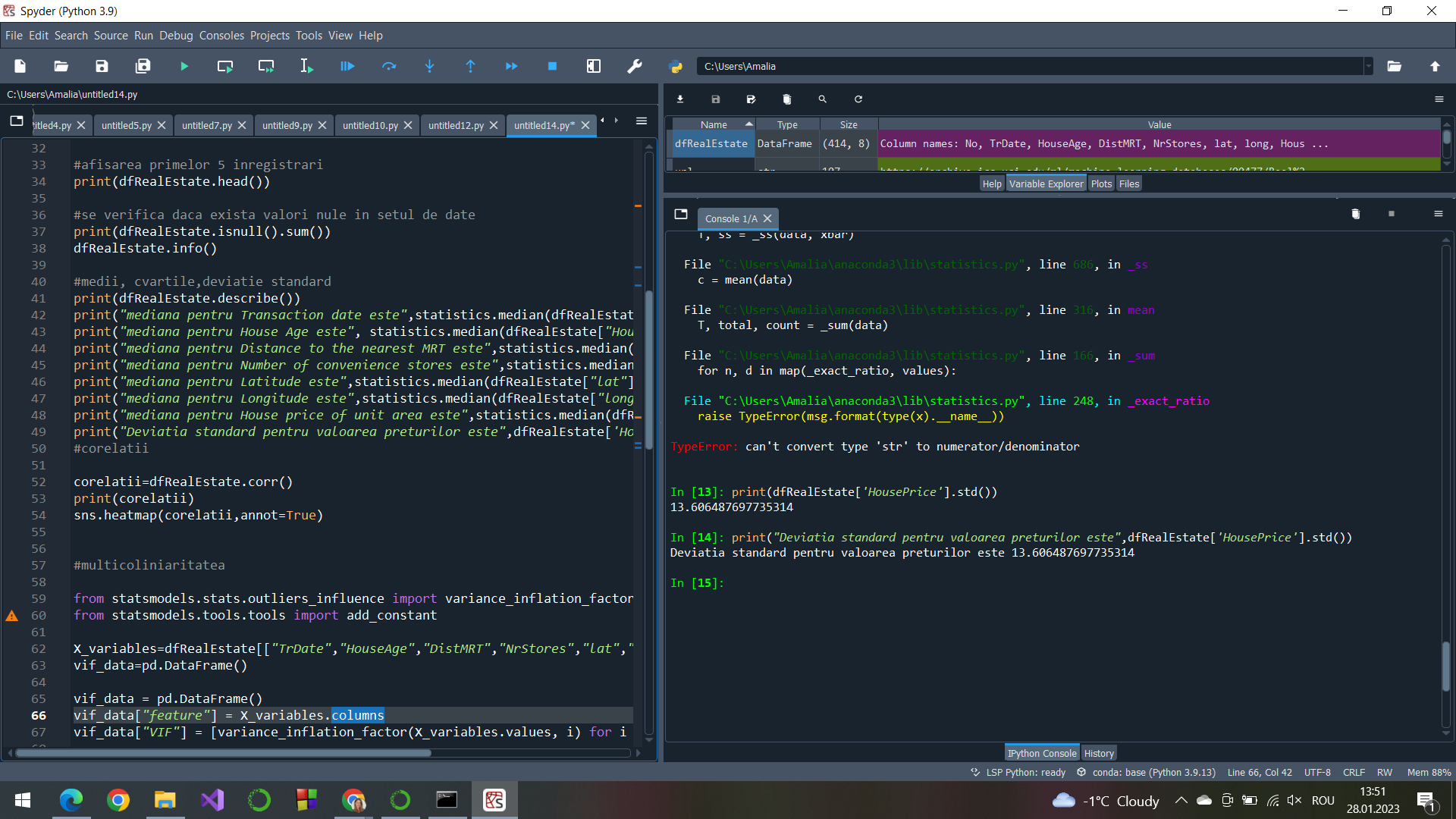


* Mediile - conform rezultatelor, în anul 2013 s-au înregistrat următoarele rezultate:
  + În medie, casele au fost vândute în luna februarie a anul 2013.
  + În medie, vechimea caselor este de 17 ani.
  + În medie, distanța față de cea mai apropiată stație de metrou este de 1 kilometru.
  + În medie, în jurul unei case se află un număr de 4 magazine.
  + În medie, casele oferite spre vânzare se află la 24° latitudine și 121° longitudine.
  + În medie, prețul unei case este de 37.98 New Taiwan Dollar/Ping.
* Cvartilele - conform rezultatelor, se cunosc următoarele:
  + prima cvartilă delimitează cele mai mici 25% din date.
  + a doua cvartilă delimitează cele mai mici 50% din date și este egală cu mediana.
  + a treia cvartilă delimitează cele mai mari 25% din date.



* Medianele - conform rezultatelor, se cunosc următoarele:



* Deviația standard - conform calculelor, se cunoaște următorul rezultat:

**4. Variabile categoriale**

* O variabilă categorială este o variabilă care poate prelua una dintr-un număr limitat și ,de obicei, fix de valori posibile.Este o variabilă care nu poate fi măsurată și nu îi poate fi calculată cantitatea.
* Într-un set de date, variabilele categoriale sunt șiruri de caractere care au însemnătate concisă când sunt citite de către un utilizator(om), însă până când nu au o valoare numerică, adică sunt transformate în dummy, calculatorul nu le poate lua în calcul pentru analiză.
* În setul de date Real Estate Value nu există variabile categoriale, deoarece toate atributele sunt formate din variabile cantitative, de exemplu “House Age”, ”distance to the nearest MRT station”, “number of convenience stores”, ”house price of unit area”, ”longitude”, ”latitude”.
* Singura excepție este “transaction date” care este o variabila care poate fi atât categorială, cât și cantitativă. În cazul în care o variabilă de tip dată ia valori, de exemplu, zilele săptămânii (luni, marti, miercuri, joi ect) sau lunile anului (ianuarie, februarie, martie etc), variabila este categorială și se ia în calcul transformarea ei în variabilă dummy, pentru a calcula regresia. În acest caz, “transaction date” este o variabilă cantitativă, deoarece reprezintă data când tranzacția a fost efectuată, adică un timp care s-a scurs. De asemenea, „transaction date” este o variabilă de tip float64, deci nu necesită să fie transformată în variabilă dummy, deoarece deja are o valoare numerică.

**5. Variabile independente**

* Variabila independentă este factorul pe care îl modificați sau îl controlați în mod intenționat pentru a vedea ce efect are.
* O variabilă dependentă este o consecință a unei variabile independente, adică este o variabilă care măsoară efectul variabilei independente asupra unităților de testare.
* Variabila independentă este reprezentată grafic pe axa x.

Cele mai importante variabile independente:

**House Age**: Vechimea casei influențează prețul deoarece, deși aceasta este poziționată într-o zonă centrală, o casă nouă este întotdeauna în topul favoritelor pentru cumpărător.

**Distance to the nearest MRT station**: Distanța parcursă de acasă până la stația de metrou este un reper important pentru cumpărător.

**Number of convenience stores**: Întotdeauna un cumpărător va prefera o zonă cu multe magazine.

**6. Corelațiile și reprezentarea grafică**

Corelația ne spune dacă două variabile sunt asociate (corelate), dar nu ne indică și relația dintre variabile.

Coeficientul de corelație r are valori cuprinse între -1 și 1.

În cazul unui coeficient de corelație pozitiv, avem o corelație directă (cele două variabile corelate variază în același sens).

În cazul unui coeficient de corelație negativ, avem o corelație inversă (când o variabilă crește, cealaltă scade).

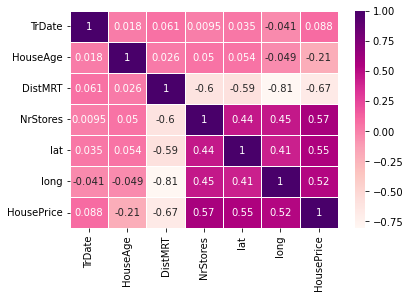
r = [0; 0,2] -corelație foarte slabă, inexistentă, nulă

r = [0,2; 0,4] - corelație slabă

r = [0,4; 0,6] - corelație rezonabilă, medie

r = [0,6; 0,8] - corelație înaltă, puternică

r = [0,8; 1] - corelație foarte înaltă (relație foarte strânsă între variabile)



*Valorile scrise în negru arată corelațiile inverse, iar cele scrise în alb corelațiile directe.*

**Corelația între variabilele independente și variabila dependentă**

* variabila dependentă: *HousePrice*
* variabilele independente:

***TrDate***

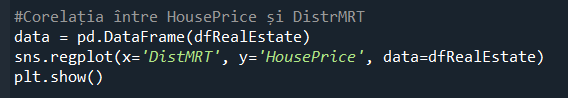
Între prețul casei și data tranzacției corelația este nulă (nu există)

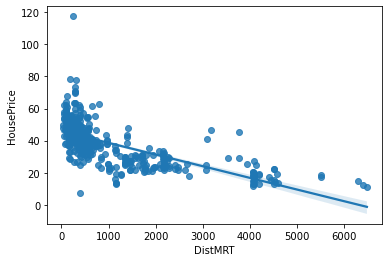
***HouseAge***

Între prețul casei și vechimea acesteia există o corelație inversă slabă (cu cât locuința este mai veche, cu atât prețul casei scade)

***DistMRT***

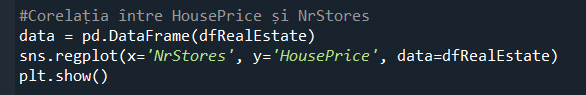
Între prețul casei și distanța până la cea mai apropiată stație de metrou există o corelație inversă puternică (cu cât distanța față de metrou este mai mare, cu atât locuința devine mai greu accesibilă și prin urmare, prețul acesteia scade)

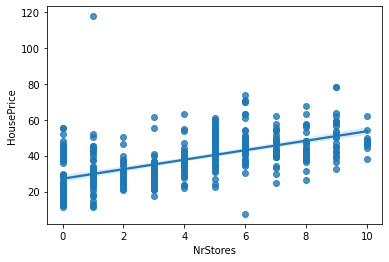




***NrStores***

Între prețul casei și numărul magazinelor din apropiere există o corelație directă medie (cu cât există mai multe magazine în apropierea locuinței, cu atât prețul acesteia crește)





***lat, long***

Între prețul casei și latitudine/ longitudine există o corelație directă medie (localizarea casei are importanță în stabilirea prețului acesteia și cu cât localizarea este mai bună, cu atât prețul casei crește)

**Corelația între variabilele independente**

*TrDate*

Observăm că între coloana “TrDate” și toate celelalte variabile independente avem o corelație foarte slabă, inexistentă.

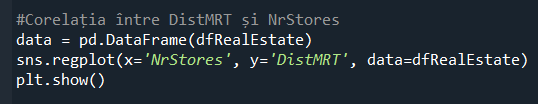
*HouseAge*

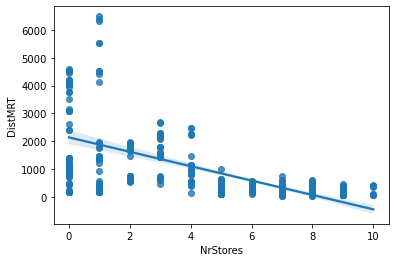
Între vechimea casei și celelalte variabile independente avem o corelație foarte slabă, inexistenă.

*DistMRT*

Între distMRT și *TrDate, HouseAge* corelația este inexistentă.

Între distanța până la metrou și *NrStores* există o corelație inversă puternică.





Între distanța până la metrou și *lat* există o corelație inversă medie.

Între distanța până la metrou și *long* există o corelație inversă foarte puternică.

*NrStores*

Între NrStores și *TrDate, HouseAge* există o corelație directă foarte slabă, inexistentă.

Între NrStores și *lat/ long* există o corelație directă medie, spre slabă.

*lat*

Între lat și *TrDate/* *HouseAge* există o corelație directă foarte slabă.

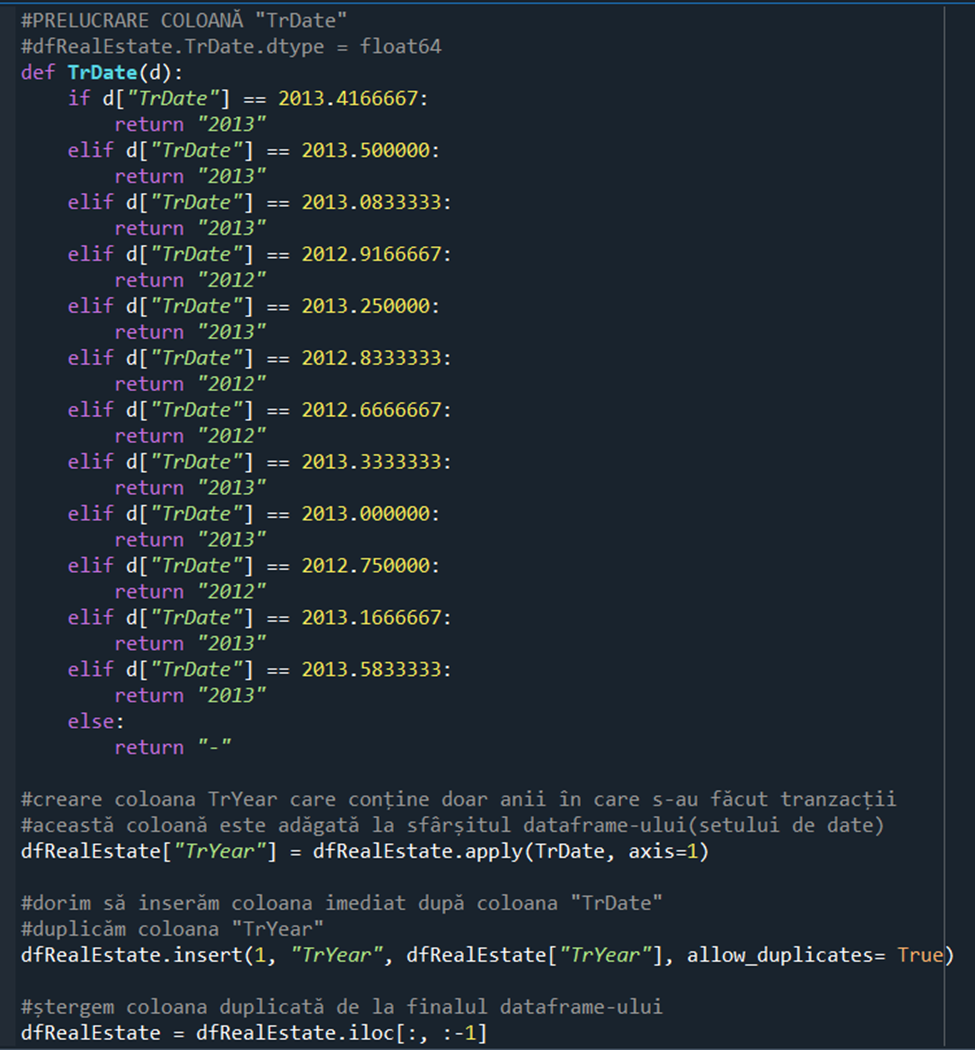
Între lat și *DistMRT/ NrStores/ long* există o corelație inversă medie.

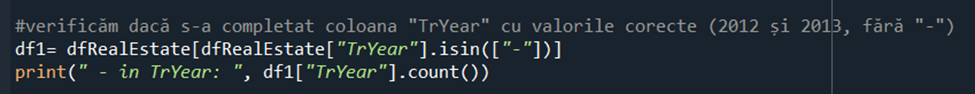
*long*

Între long și *TrDate/ HouseAge* există o corelație directă foarte slabă.

Între long și *DistMRT* există o corelație inversă foarte puternică.

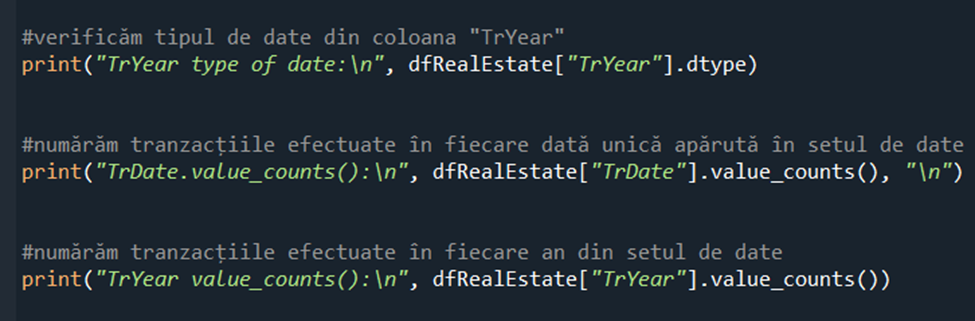
Între long și *NrStores, lat* există o corelație directă medie spre slabă.

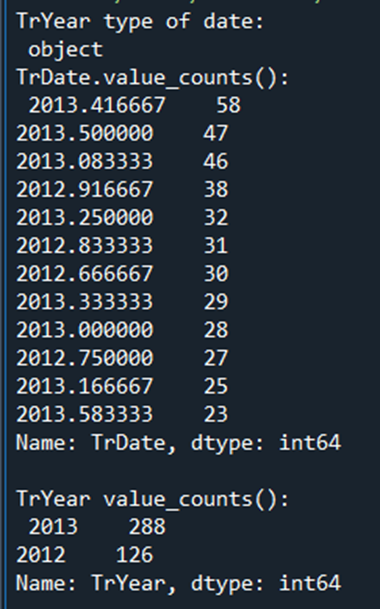




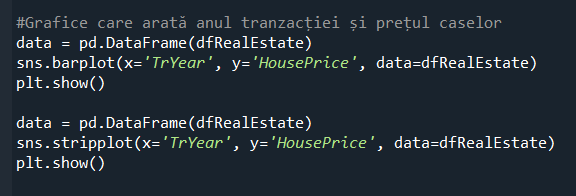
Consolă:

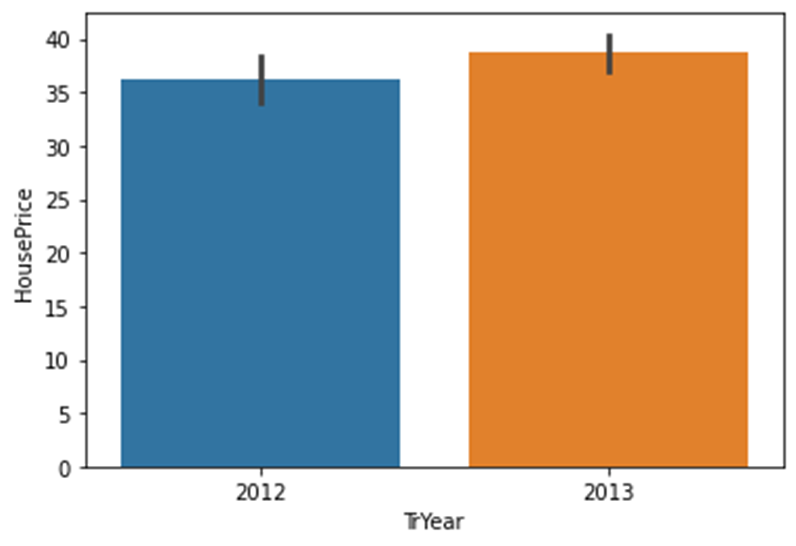


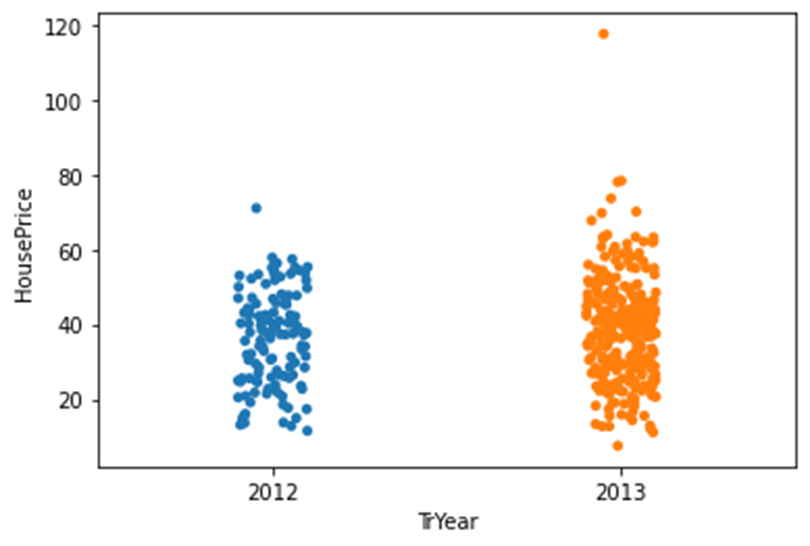




Dacă transformăm tipul de date din coloana TrYear în datetime apărea la TrYear.value\_counts() 2013-01-01/ 2012-01-01, ceea nu era tocmai corect.

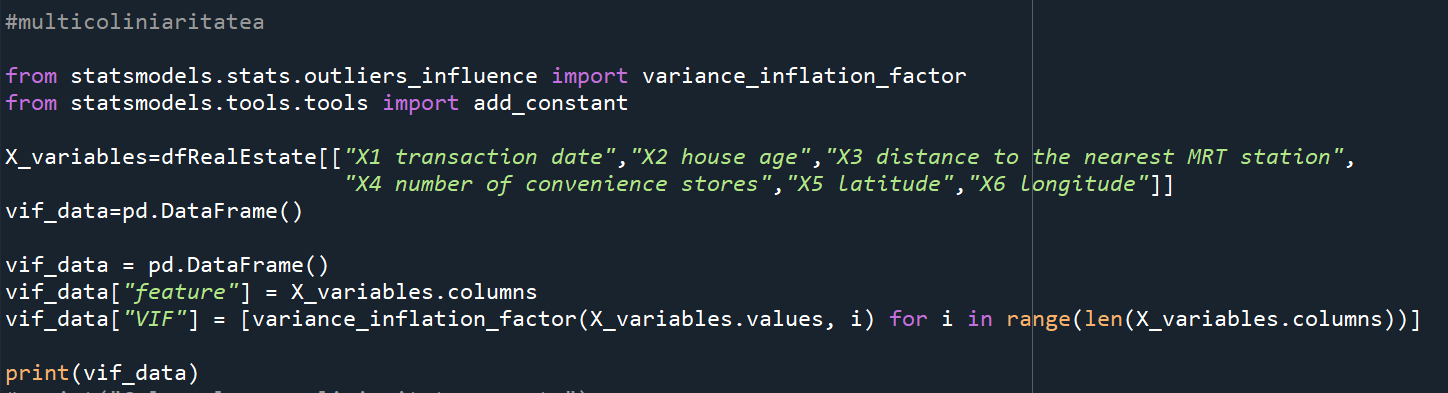




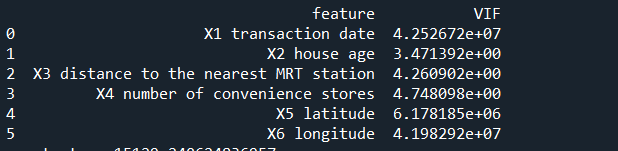


**7. Multicoliniaritatea**

* Multicoliniaritatea într-un model de regresie multiplă apare atunci când un grup de variabile independente sunt puternic corelate între ele. În acest caz, prin includerea în model a unei variabile din grup, restul variabilelor din grup nu mai aduc o informaţie semnificativă. Simultan are loc o supraevaluare a coeficientului de determinare ceea ce poate denatura interpretarea modelului.
* Multicoliniaritatea este detectată folosind două metode, prima fiind calcularea toleranței variabilei xi, iar a doua metodă este calcularea factorului de inflație a variației prin formula .
* Dacă valoare VIF<=5, atunci coliniaritatea nu este puternică, iar dacă VIF>5, se denotă o coliniaritate puternică. Am folosit metoda factorului de inflație a variației, pentru a determina dacă exista multicoliniaritate în setul nostru de date.

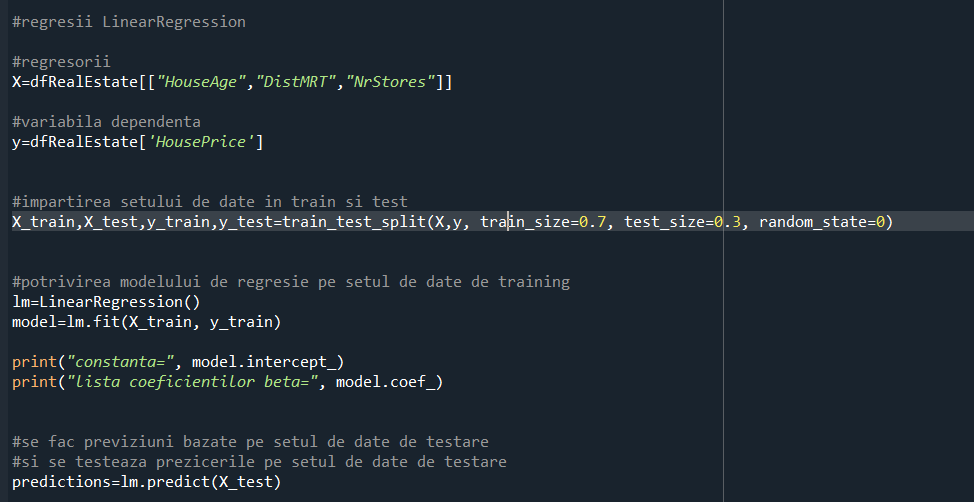


* În rezultat, am obținut pentru atributul “X5 latitude” valoarea VIF>5,respectiv 6,17. Astfel,pentru calcularea regresiei liniare vom exclude din setul de date aceasta variabilă si variabila “long”, deoarece aceste variabile sunt puternic corelate.

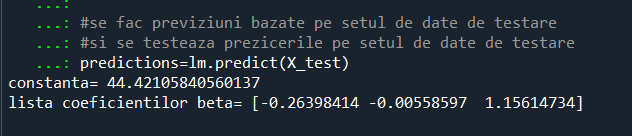


**8. Regresii**

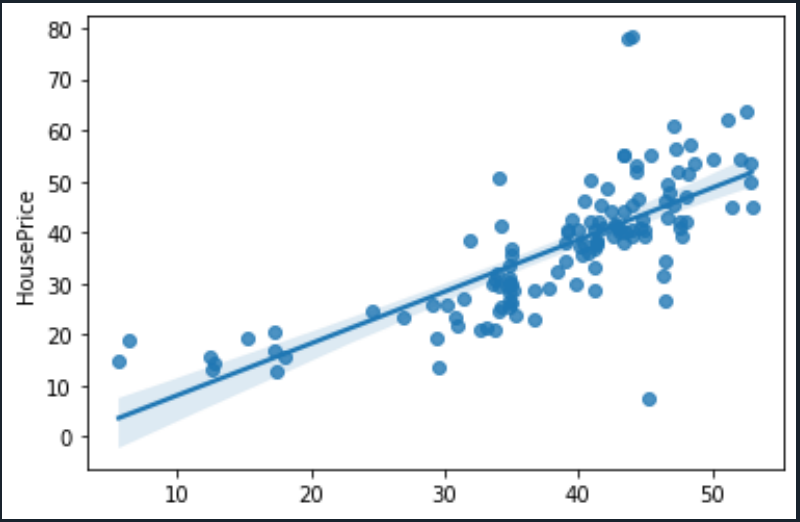
* **Pentru setul de date Real Estate Value am implementat modelul Regresiei Liniare din pachetul sklearn.linear\_model.**



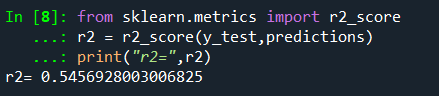
* În urma rulării programului, am obținut următoarele rezultate pentru intercept și constantele beta.



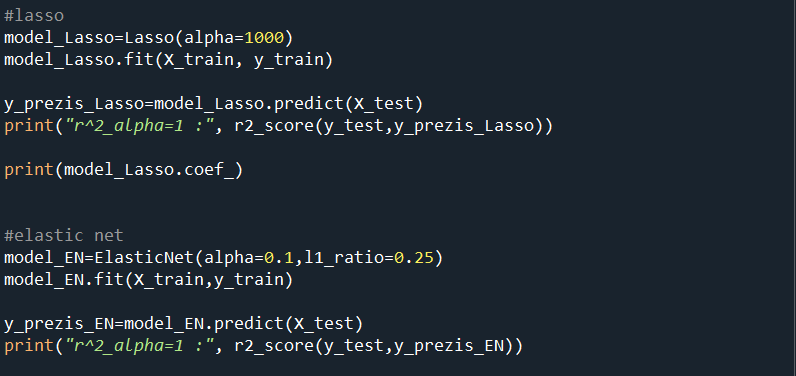
* Astfel, prețul pentru un ping (3.3 metri patrati) în dolari New Taiwan ar fi egal cu 44,42 dolari, atunci când restul variabilelor independente ar fi egale cu 0.
* Coeficienții beta arată gradul de modificare a variabilei dependente atunci când variabila independentă se modifică cu 1 unitate. Dacă vârsta casei s-ar mari cu un an, atunci prețul pentru un ping(3.3 metri pătrați) ar scădea cu 2.26 dolari taiwanezi.Dacă distanța până la cea mai apropiată stație de metrou ar crește cu 1 metru, atunci prețul unui ping ar scădea cu 0.005 dolari.



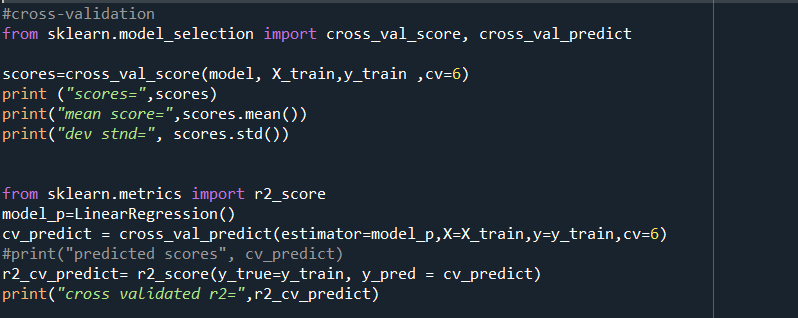
* În imagine, este reprezentat graficul Regresiei liniare. Dreapta reprezintă valorile prezise ale HousePrice, iar punctele de pe grafic sunt valorile de testare(valorile reale). Distanța de la axă până la punct reprezintă eroarea de calcul.
* Coeficientul de determinare:



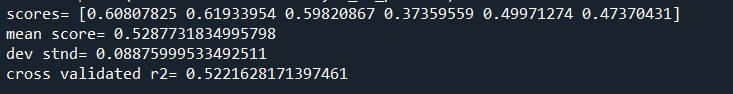
* Variabila dependentă **HousePrice** depinde în proporție de 0.54 față de restul variabilelor independente, fiind o legătură medie. 54% din variația variabilei HousePrice este explicată de variația celorlalte variabile.



* **Cross-validation.**



* **Rezultate cross validation.**



Pentru a minimiza riscul de overfitting sau underfitting al modelului de regresie liniare, am utilizat cross-validation. Astfel, coeficientul de determinare a fost normalizat la valoarea de 0.52.

* Regresia Lasso și Ridge sunt câteva dintre tehnicile de reducere a complexității modelului și de a preveni overfitting prin penalizarea coeficienților predictorilor care pot exista datorită unui model liniar simplu.

