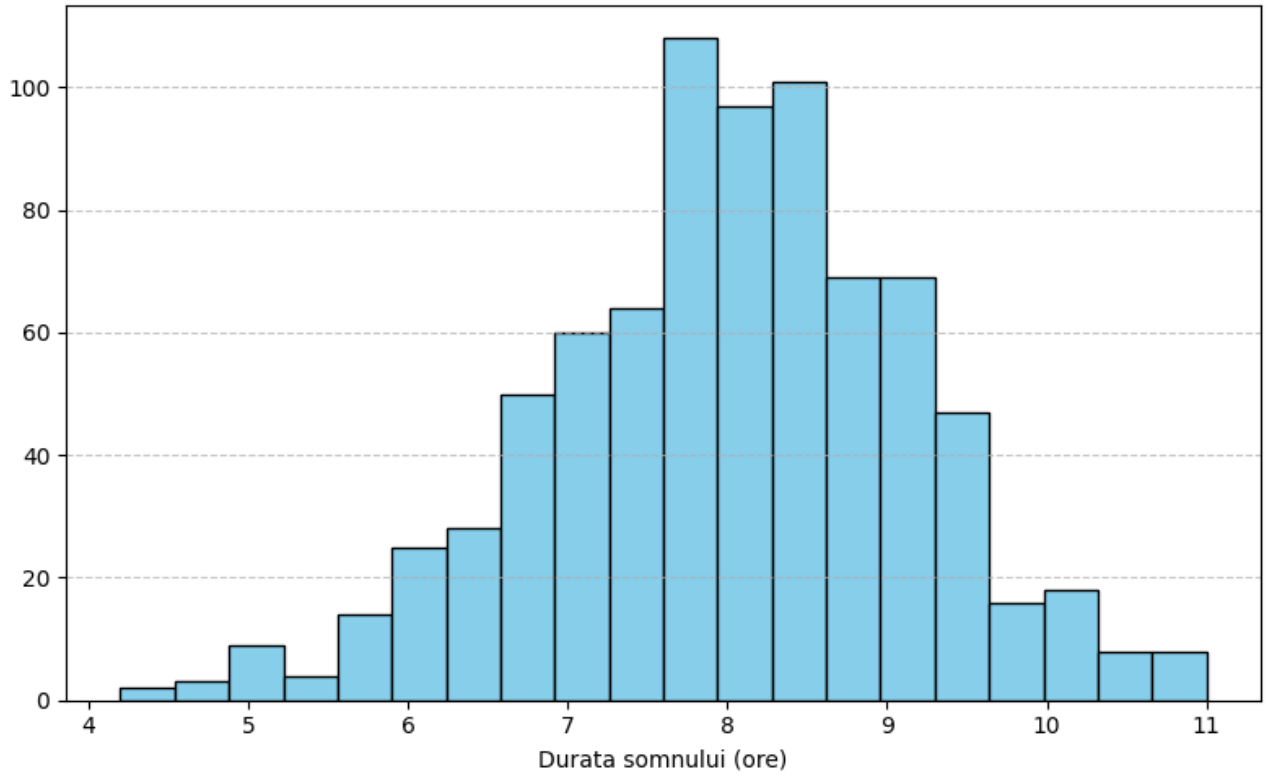
Am ales o problema de regresie liniara :predicția duratei de somn în funcție de obiceiuri zilnice. Am optat pentru generarea sintetică a unui dataset care să aibă sens contextual. Setul de date este compus din 9 coloane: varsta (numar intreg), gen ('M' sau 'F'), ore petrecute in fata unui ecran (nr real), numarul de cafele consumate (numar intreg), timpul de sport facut (masurat in minute - nr real), nivelul de stres (scazut/mediu/ridicat), ora de culcare (nr real; ora este trata ca un intreg - 0.1h = 6min) si coloana finala - durata de somn, pe care vom aplica si regresia liniara. Acesta este calculata cu o formula intuitiva, de la 8 ore, scadem proportional cu nr de ore de ecran, nr de cafele, zgomotului, si creste cu practicarea sportului.



Putem observa ca somnul are distributia normala in jurul 7.5-8h.

VALORI LIPSĂ - X_TRAIN:
VARSTA 22
GEN 13
ORE_ECRAN 22
CAFEA 16
MINUTE_SPORT 24
NIVEL_STRES 13
ORA_CULCARE 18
ZGOMOT 16
DTYPE: INT64

VALORI LIPSĂ - X_TEST:
VARSTA 3
GEN 1
ORE_ECRAN 4
CAFEA 8
MINUTE_SPORT 6
NIVEL_STRES 2
ORA_CULCARE 7
ZGOMOT 5

Valorile lipsa au fost alese random din cele 8 coloane, 2.5% din numarul total de valori. De asemenea ele au fost inlocuite cu media din acea coloana.

X_TRAIN:

	VARSTA	ORE_ECRAN	CAFEA	MINUTE_SPORT	ORA_CULCARE
COUNT	640.000000	640.000000	640.000000	640.000000	640.000000
MEAN	39.082870	5.176179	1.418264	30.192545	16.969311
STD	11.990449	1.966347	1.190569	9.458536	9.266774
MIN	18.000000	0.000000	0.000000	-0.195122	0.000000
25%	29.000000	3.800000	1.000000	24.349735	17.286600
50%	40.000000	5.229764	1.000000	30.124952	21.900000
75%	49.250000	6.425000	2.000000	36.609153	23.000000
MAX	59.000000	10.200000	6.000000	61.931076	24.000000

	ZGOMOT
COUNT	640.000000
MEAN	69.659584
STD	5.599490
MIN	60.004820
25%	64.826467
50%	69.797718
75%	73.996161
MAX	79.978094

X_TEST:

	VARSTA	ORE_ECRAN	CAFEA	MINUTE_SPORT	ORA_CULCARE
COUNT	160.000000	160.000000	160.000000	160.000000	160.000000
MEAN	36.994634	5.283869	1.664653	30.057360	17.598789
STD	11.526640	1.940803	1.191366	10.229151	8.852597
MIN	18.000000	0.000000	0.000000	3.957861	0.000000
25%	28.000000	4.100000	1.000000	23.534441	17.286600
50%	37.000000	5.214882	1.543058	30.124952	22.000000
75%	46.250000	6.400000	2.000000	36.191204	23.100000
MAX	58.000000	9.900000	6.000000	61.377485	24.000000

	ZGOMOT
COUNT	160.000000
MEAN	69.983189
STD	5.678692
MIN	60.008031
25%	65.244598
50%	69.821907
75%	74.592267
MAX	79.948459

X_TRAIN:

	GEN	NIVEL_STRES
COUNT	627	627
UNIQUE	2	3
TOP	M	MEDIU
FREQ	328	313

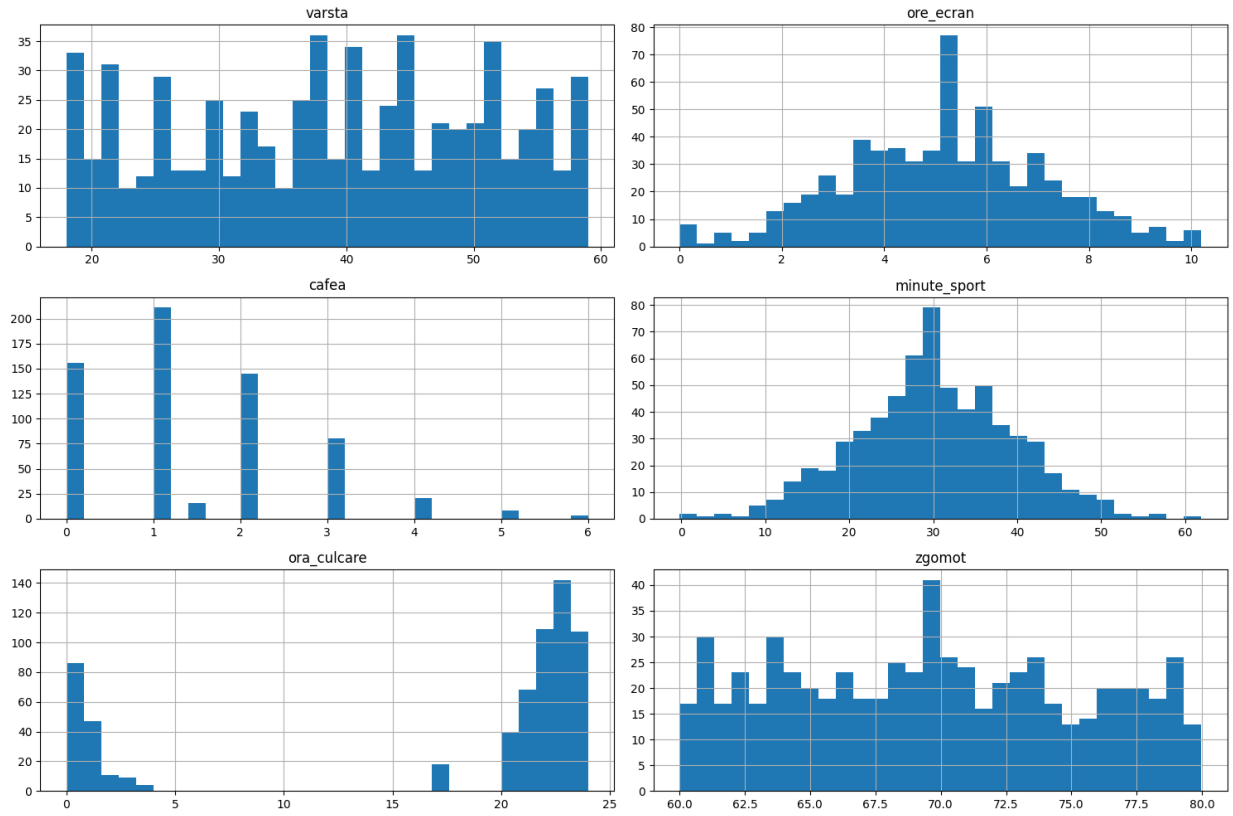
X_TEST:

	GEN	NIVEL_STRES
COUNT	159	158
UNIQUE	2	3

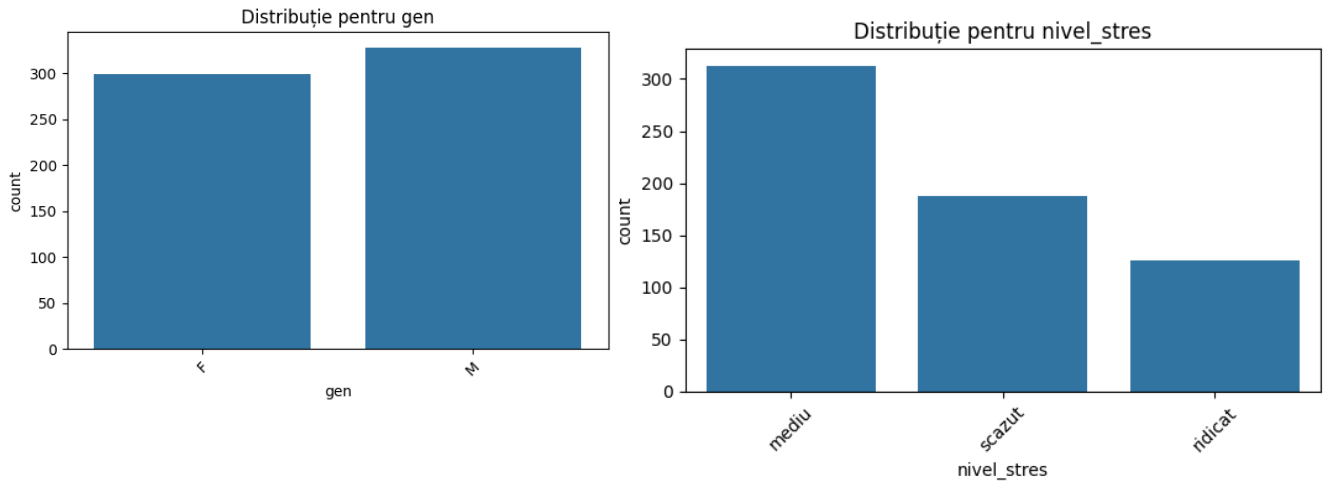
TOP	M	MEDIU
FREQ	80	76

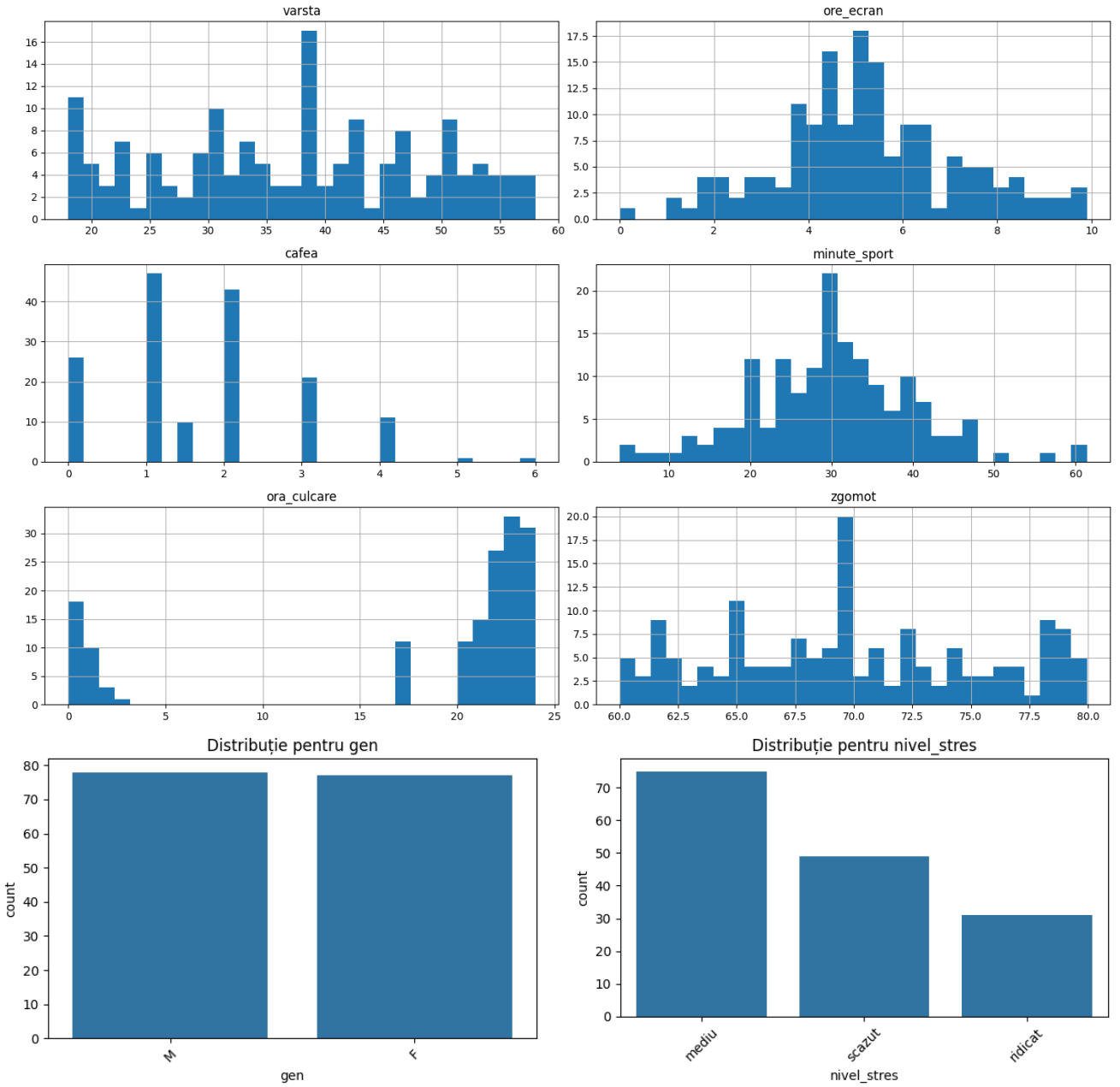
Pentru fiecare set de date, le-am analizat cu functia “.describe” avand informatii pentru fiecare coloana precum: numarul, media, standard devietion, min, max, si cum se incadereaza datele, rescativ la cele care nu sunt numerice, numarul de variante, si frecventa acestora.

Mai jos fac graficele pentru fiecare variabila, atat pentru X_train cat si pentru X_test, acestea vor semana, dat fiind distribuirea variabilelor, graficele trebuie sa fie similare, cun mentiunea unor diferente, dat fiind nr mai mic de entitati De asemea se poate vedea cum in mijlocul grafului sunt mult mai multe date, din cauza faptului ca valorile lipsa au fost inlocuite cu media coloanei respective.

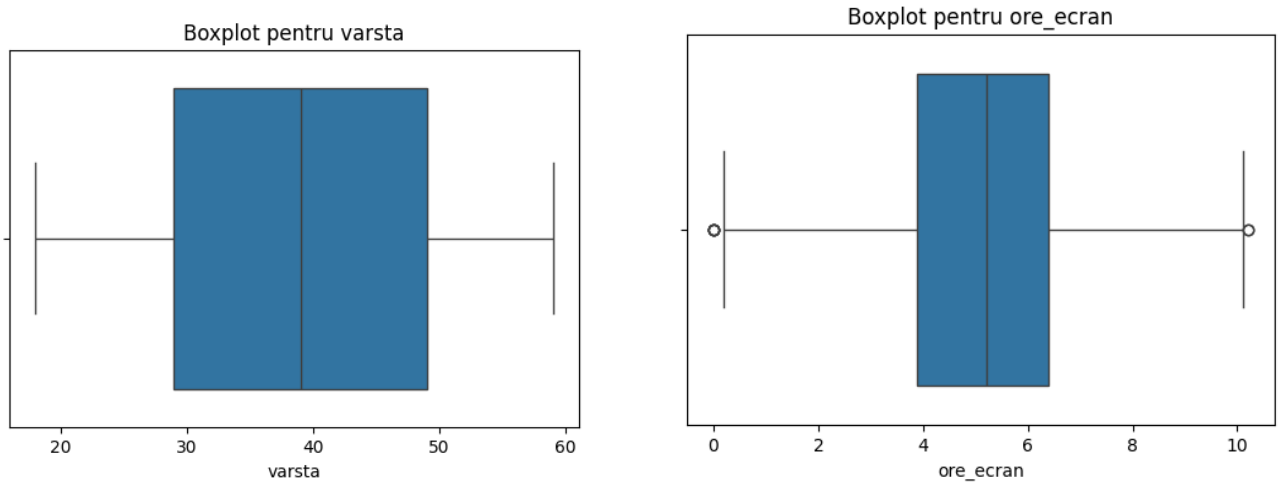


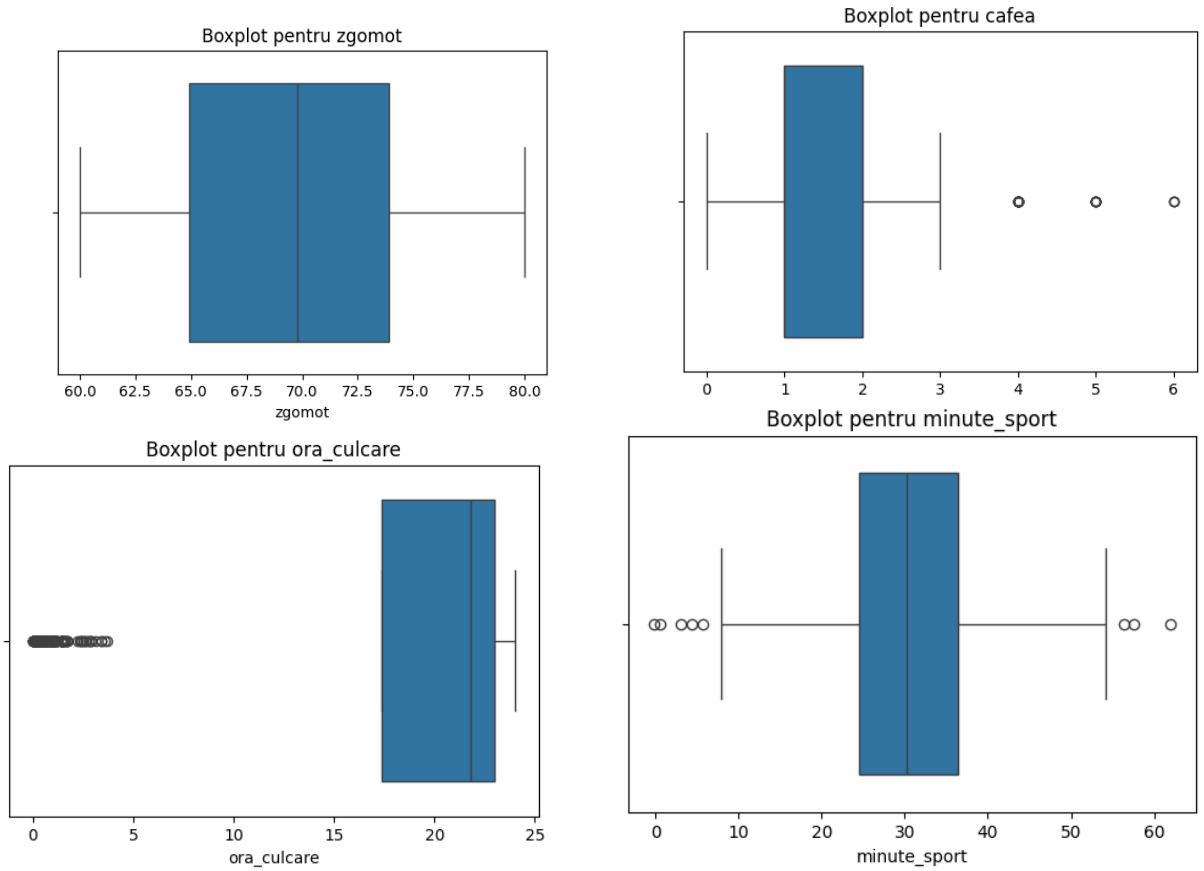
Se poate observa cum variabile precum varsa si zgomotul care au o distributie random, nu au un grafic uniform, pe cand orele pe ecran, ora de culcare si minutele sport au o distributie normala, iar numarul de cafele sunt distribuite tip Poisson. De asemena se vede la fiecare grafic un salt mare in dreptul valorii medii din cauza datelor lipsa care au fost inlocuite cu acestea.
Pentru nivel stres si sex au fost o distributie random.



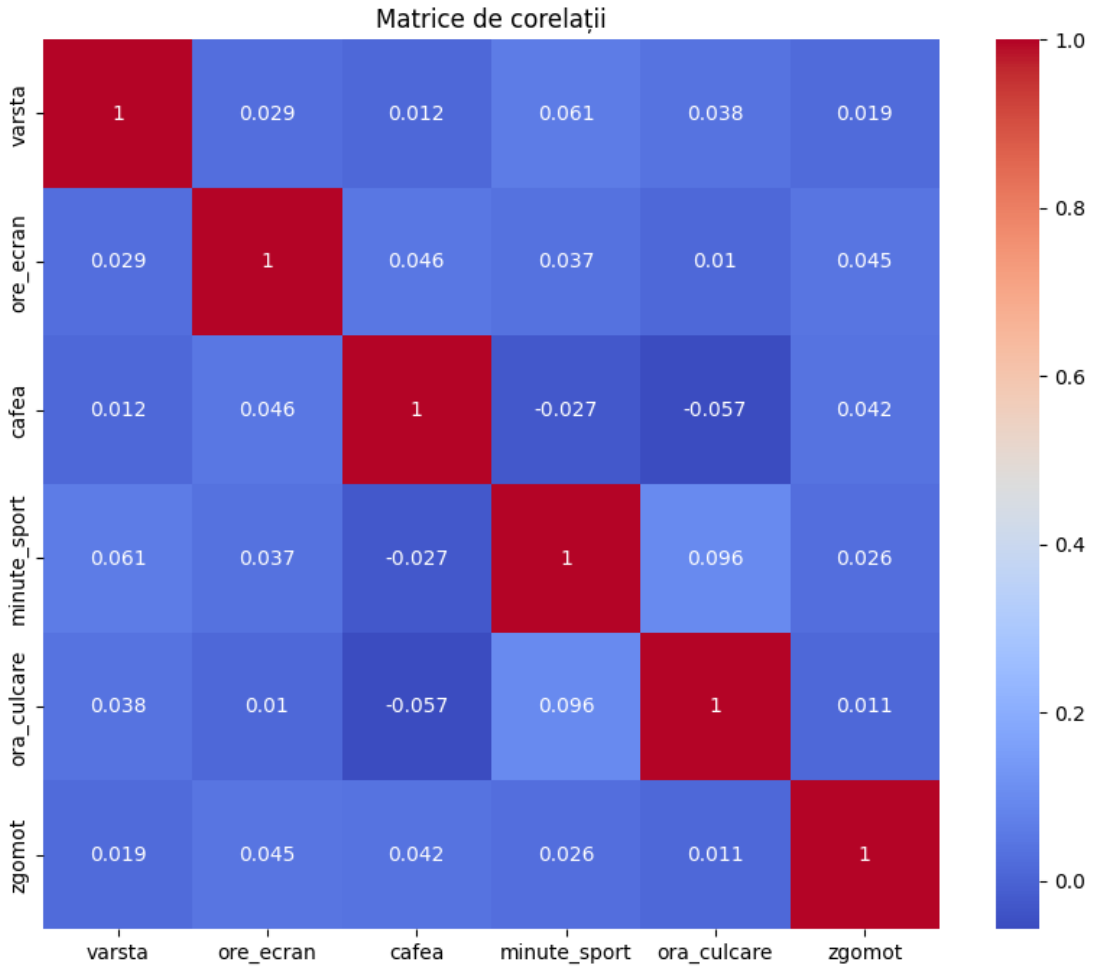


De asemenea, **toate** graficele sunt similare si pentru setul de test, avand in vedere ca distrubuirea in cele 2 seturi a fost random.

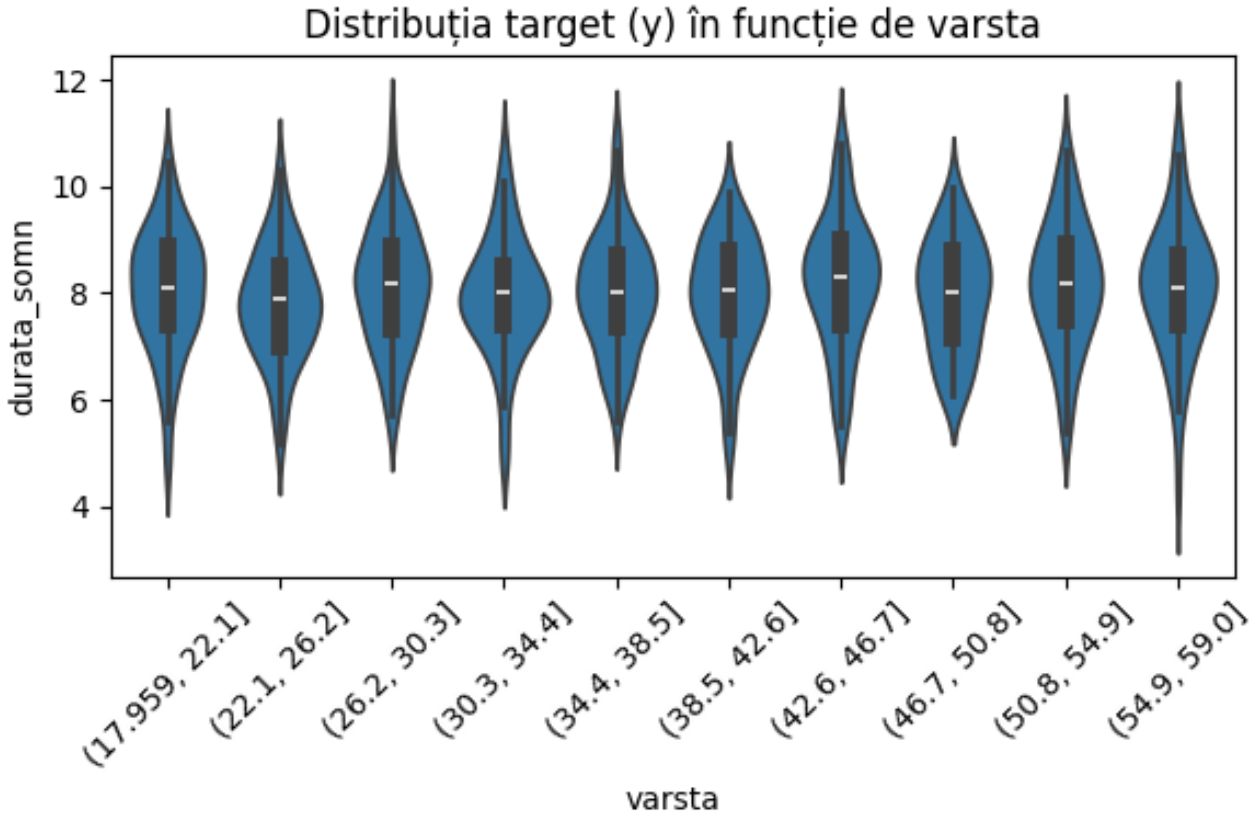


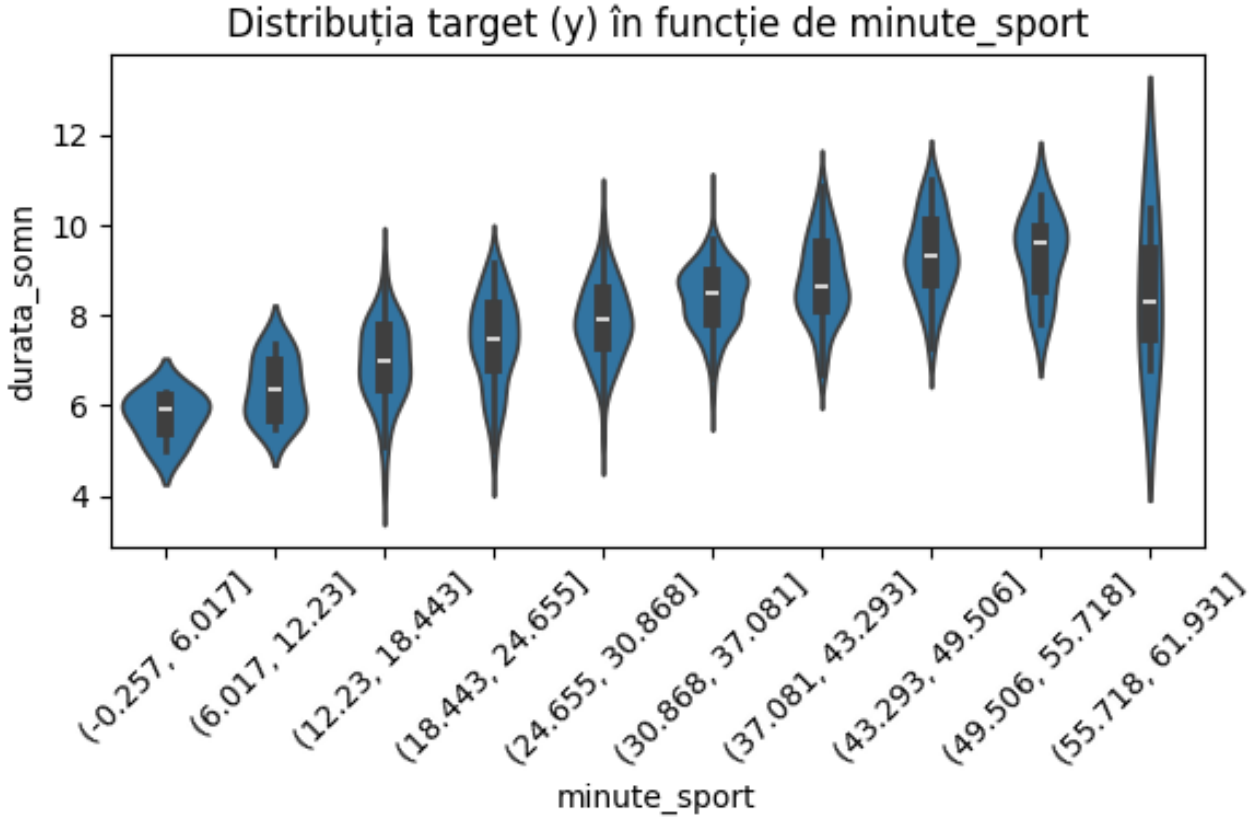


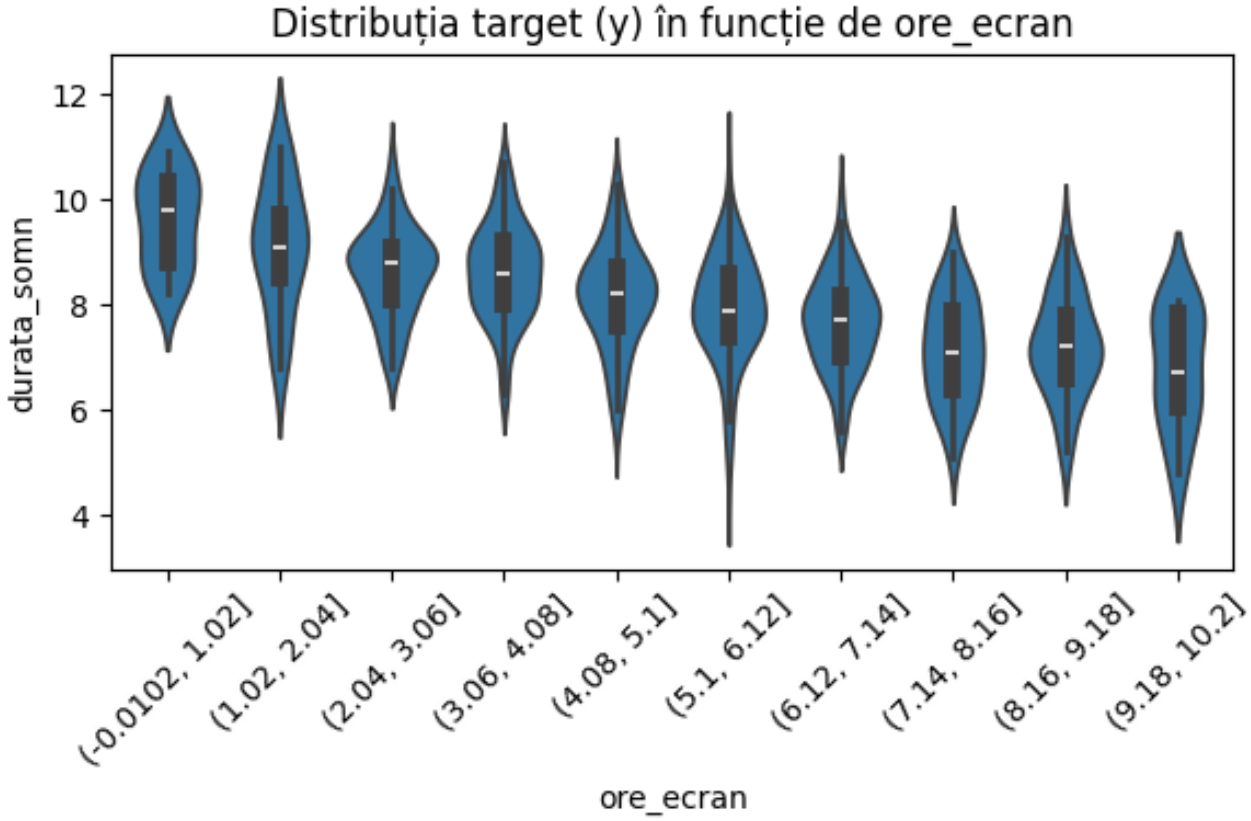
Se poate vedea cum la distrubirea random nu prea exista outlineri, dar range-ul este mare. Pentru distrubuirea Poisson se observa outlire-ri mai multi, la fel ca si la minutele de somn. La ora de culcare, fraficul nu este reprezentativ, din pacate, intrucat ora 24=0, dar acesta nu ia in calcul faptul prezentat si indica o distribuire a valorilor incorecta.

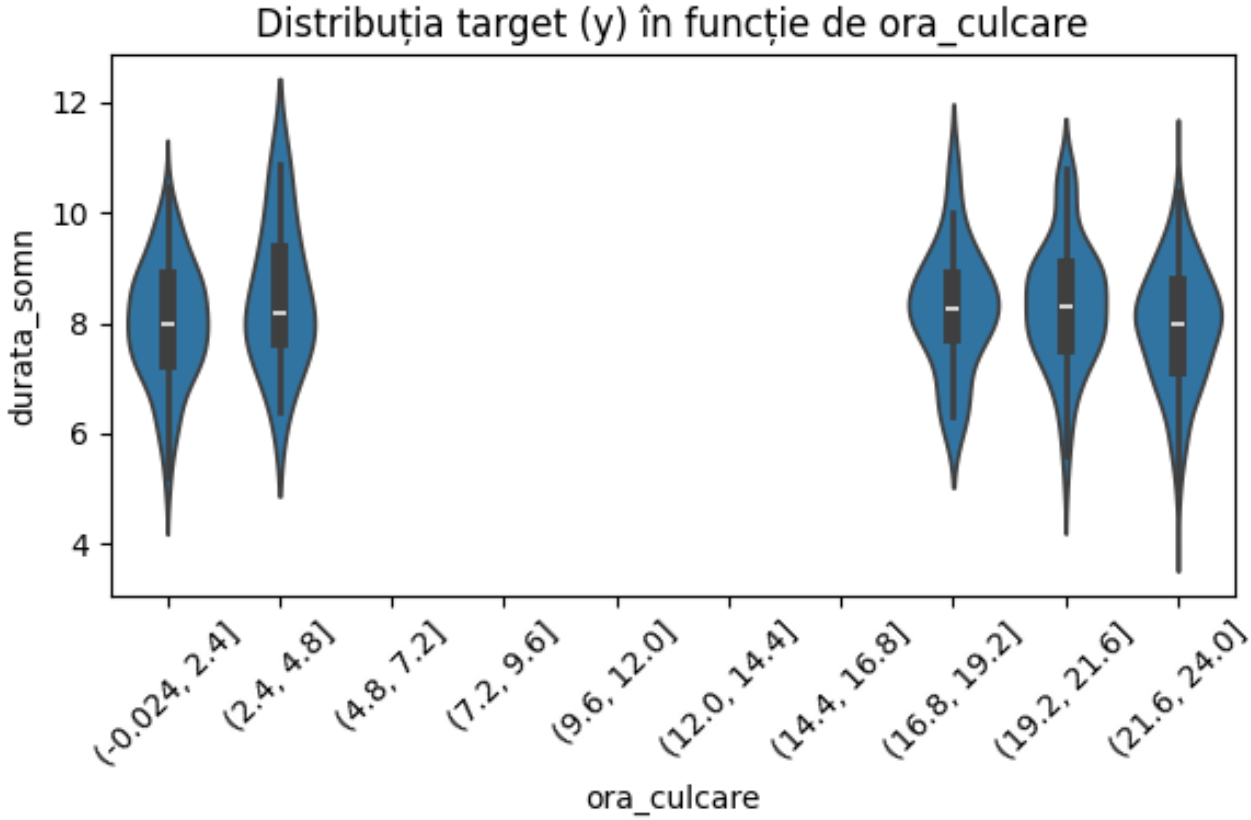


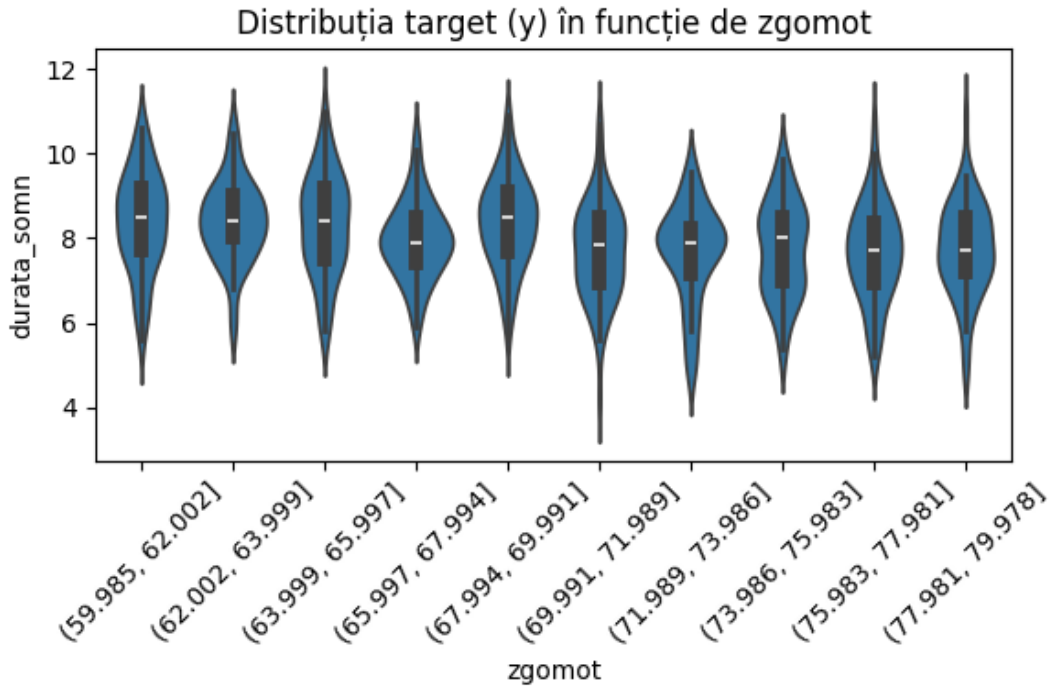
Matricea de corelatie evidentiaza faptul ca toate varibilele au fost generate random independent, neavand corelatii mai mari |cor| < 0.96

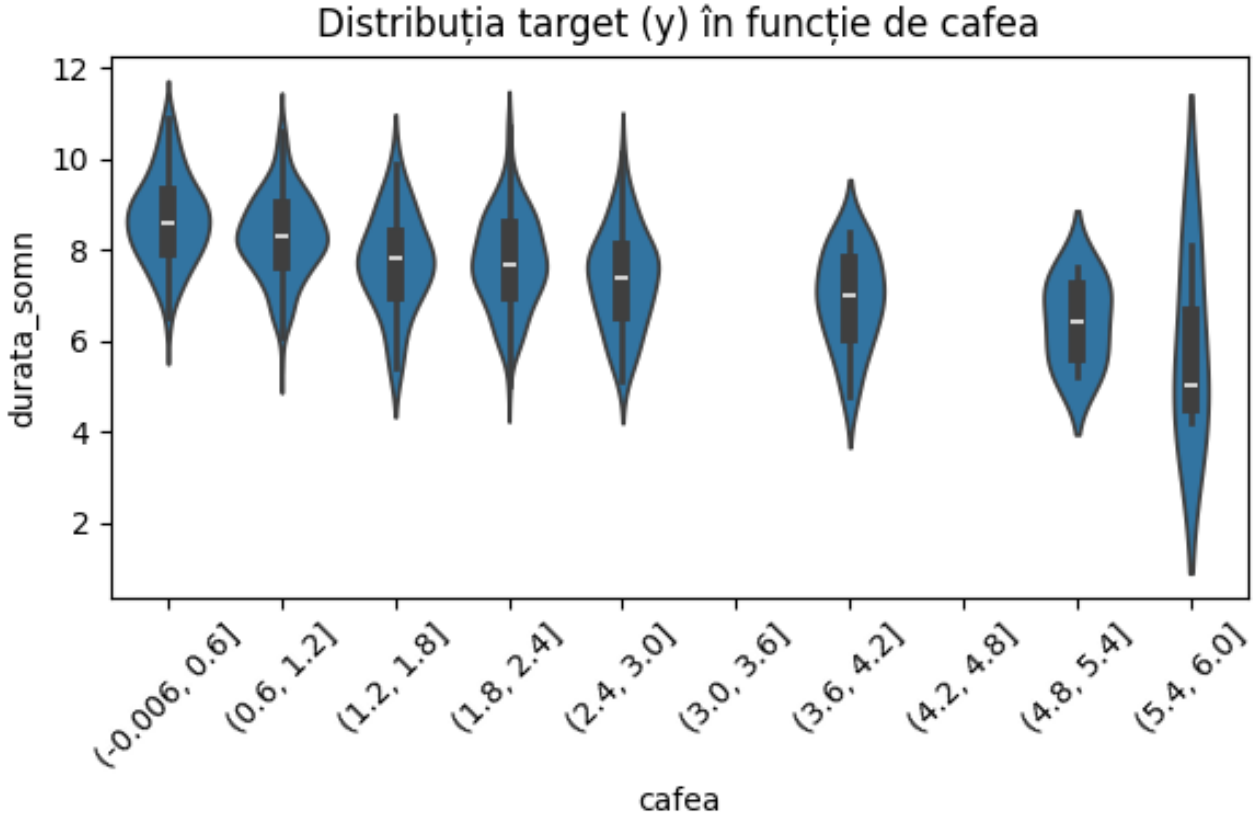





Distribuția target (y) în funcție de ore_ecran
durata_somn
12
10
8
6
4
(-0.0102, 1.02]
(1.02, 2.04]
(2.04, 3.06]
(3.06, 4.08]
(4.08, 5.1]
(5.1, 6.12]
(6.12, 7.14]
(7.14, 8.16]
(8.16, 9.18]
(9.18, 10.2]
ore_ecran


Distribuția target (y) în funcție de ora_culcare
durata_somn
12
10
8
6
4
(-0.024, 2.4]
(2.4, 4.8]
(4.8, 7.2]
(7.2, 9.6]
(9.6, 12.0]
(12.0, 14.4]
(14.4, 16.8]
(16.8, 19.2]
(19.2, 21.6]
(21.6, 24.0]
ora_culcare

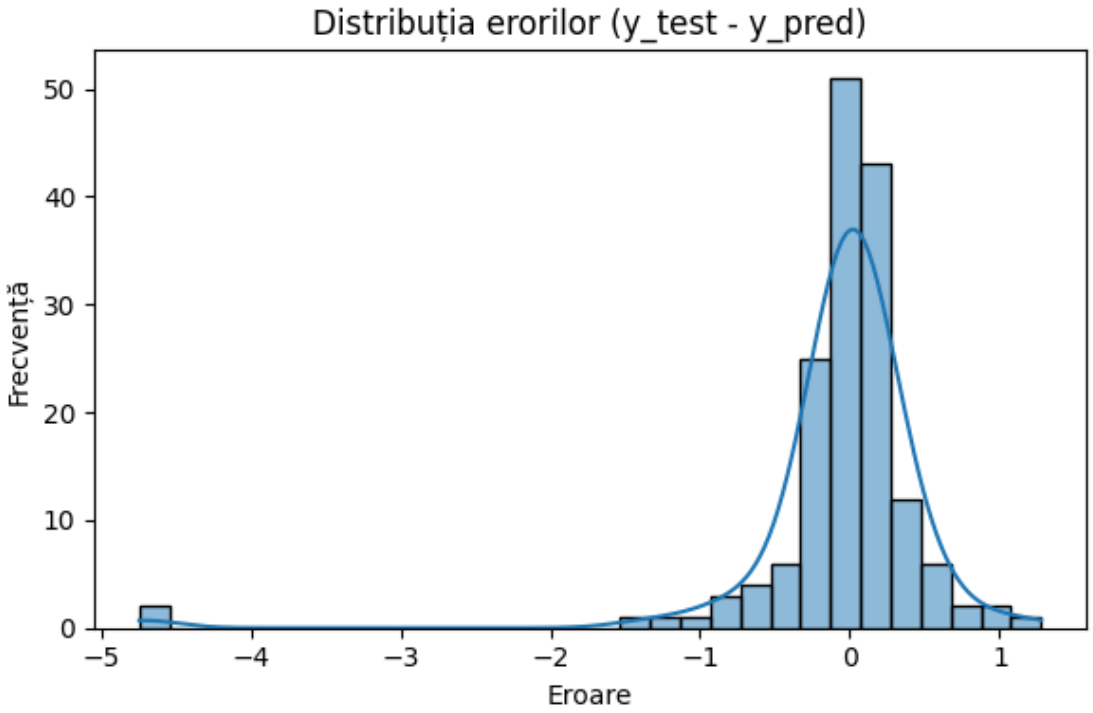




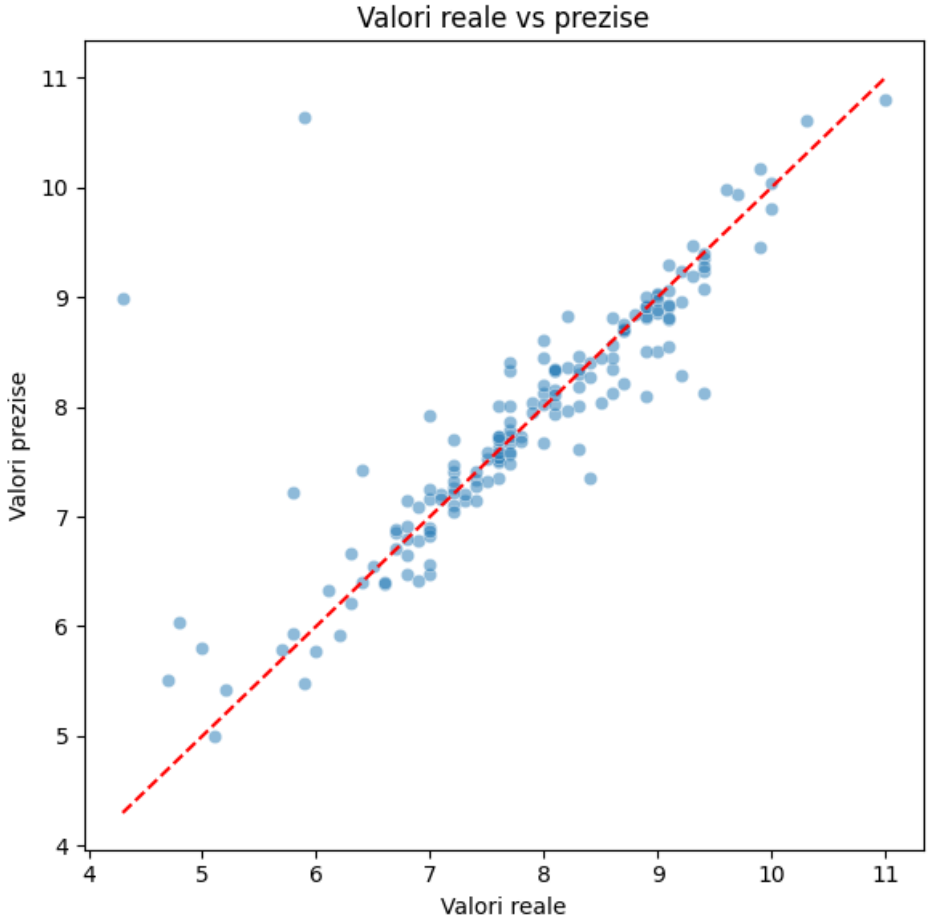
Privitor la analiza rezultatelor cu varibilele tinta, de poate observa corelatia pe care am dat-o in formula privitoare la fiecare coloana: durata somnului este calculata cu o formula intuitiva, de la 8 ore, scadem proportional cu nr de ore de ecran, nr de cafele, zgomotului, si creste cu practicarea sportului.

RMSE: 0.63
MAE: 0.29
R^2: 0.72

Modelul de regresie folosit în proiect a avut performanțe bune, cu un scor R^2 de 0.72, ceea ce înseamnă că explică aproximativ 72% din variația datelor. Erorile nu au fost foarte mari – RMSE a fost 0.63, iar MAE 0.29 – deci predicțiile au fost destul de apropiate de valorile reale



Din graficul de distributie a erorilor se poate vedea cum antrenarea modelului a fost cu succes, intrucat erorile sunt in mare parte mici.



Nu in ultimul rand, se poate observa din graficul de mai sus, ca modelul este eficient, avand o precizie buna, cu mici abateri la valori extreme.

Pentru o mai buna colaborare am folosit GitHub pentru a stoca proiectul si a putea rula si configura in google colab.
https://github.com/alexf05/pclp3.git
https://colab.research.google.com/github/alexf05/pclp3/blob/main/TEMA.ipynb