Sandu Bogdan-Stefan(311CC)

 # PCLP3-Proiect

Am decis sa implementez un model de invatare automata ce prelucreaza un set de date ale unor pacienti suspectati de cancer. Printre datele din set se numara varsta, cativa parametri fizici generali, un set de analize de sange, precum si 11 markeri tumorali, ce reprezinta dimensiunile si parametrii formatiunii ce ii face suspecti pe pacienti(aceasta formatiune poate fi benigna(chist) sau maligna(tumora...)). In functie de toti acesti parametri, printre care se numara intregi, numere reale si siruri de caractere, voi incerca sa determin daca pacientul sufera sau nu de boala.

Aspecte cu privire la implementare se vor regasi atat in README, cat si in cod. De asemenea, tot aici voi adauga si comentariile si observatiile cu privire la graficele din folder-ul atasat.

Pentru inceput, vreau sa spun ca am gasit un dataset "curat", fara valori lipsa sau zgomot (valori aberante), dar tot am facut verificarile necesare, pentru a proba acest lucru. Dupa citirea dataset-ului si verificarea valorilor neconforme, am impartit datele 80/20 si am crearet fisierelor csv pentru train si test. Ulterior, dupa afisarea dimensiunilor acestora, am afisat statisticile pentru ambele seturi, atat pentru datele numerice, cat si pentru cele "categorice".

Acum, despre grafice: In principiu, cum se reflecta din boxplot-uri, nu avem outliers decat la BMI, ceea ce inseamna ca datele sunt grupate si nu avem discrepante foarte mari, De asemenea, si din distributii se observa acelasi lucru, deoarece graficul este, si el, oarecum liniar. In ciuda acestui fapt, care, in teorie, este benefic pentru procesul de invatare, un lucru decisiv ce intervine la precizia relativ redusa a modelului (0., adica probabilitate de 60% de a produce rezuktatul corect) se poate observa din matricea de corelatie,

unde putem observa ca doar BMI(Body Mass Index), greutatea si inaltimea au o legatura stransa, lucru care este

si logic, deoarece BMI-ul este calculat in functie de cele doua. Pentru restul, dependentele sunt infime, ceea ce,

cred eu, deruteaza modelul, asa ca as spune ca are loc un mic "underfitting". De asemenea, as mai pune acest lucru

si pe seama faptului ca, in dataset-ul de antrenament, datele de pe coloana target, "Has_Target", nu sunt repartizate

chiar in mod egal, ceea ce conduce iarasi catre un dezechilibru in favoarea valorii predominante, "No"(consulta

"output.txt"). Un lucru interesant este faptul ca, atunci cand adaug in functia "LogisticRegression" parametrul

"class_weight='balanced'" precizia, scade de la 0.61 la 0.6, dar am decis sa ma folosesc de acesta deoarece, desi, aparent

produce un rezultat mai slab, acest lucru e inselator, pentru ca trebuie sa ne uitam si la ceilalti doi indici, "recall"

si "f1-score", care, pentru producerea rezultatului "Yes" erau neglijabili in primul caz, din cauza dezechilibrului.

Teoretic, in medicina, este de dorit un recall mare, pentru ca e de preferat sa gasesti un caz fals pozitiv decat sa ratezi

un caz real .Cum se vede si pe statistica, pentru cazul in care un pacient este sanatos, este o sansa mai buna ca modelul sa dea

rezultatul corect decat pentru un pacient bolnav, pentru care un rezultat este corect in 52% din cazuri(practic,

da cu banul). De asemenea, cred ca datasetul au o parte din "vina", deoarece corelatia dintre istoricul

aparitiei bolii in familie si diagnostic ar trebui sa fie, din nou, mai stransa. Toate cele de mai sus sunt reflectate in

matrciea de confuzie, unde se poate vedea clar, ca modelul are dificultati in depistarea cazurilor pozitive.

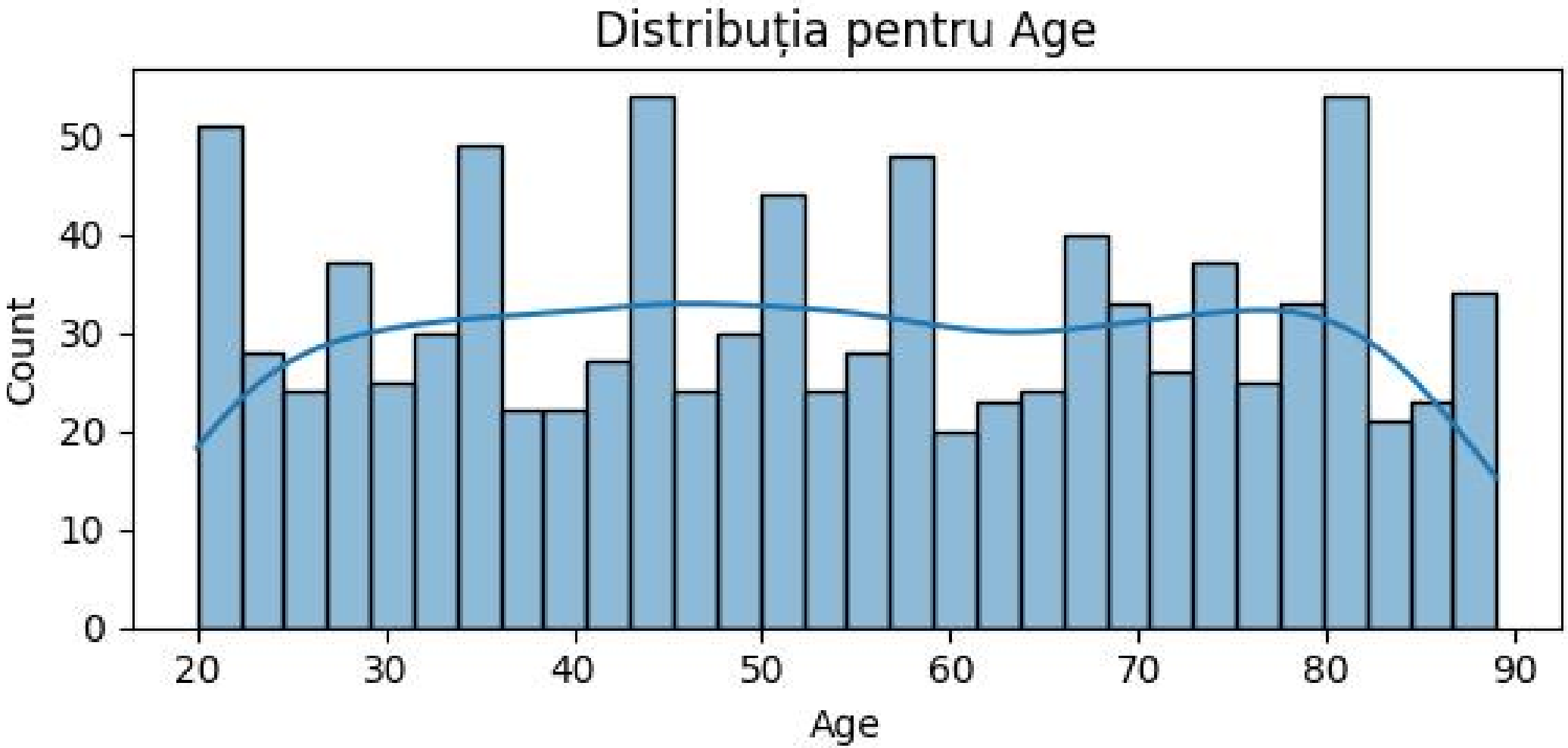
Poate ca ar fi fost mai potrivit sa predomine rezultatele "Yes"(din motivul de fals pozitiv de mai sus).

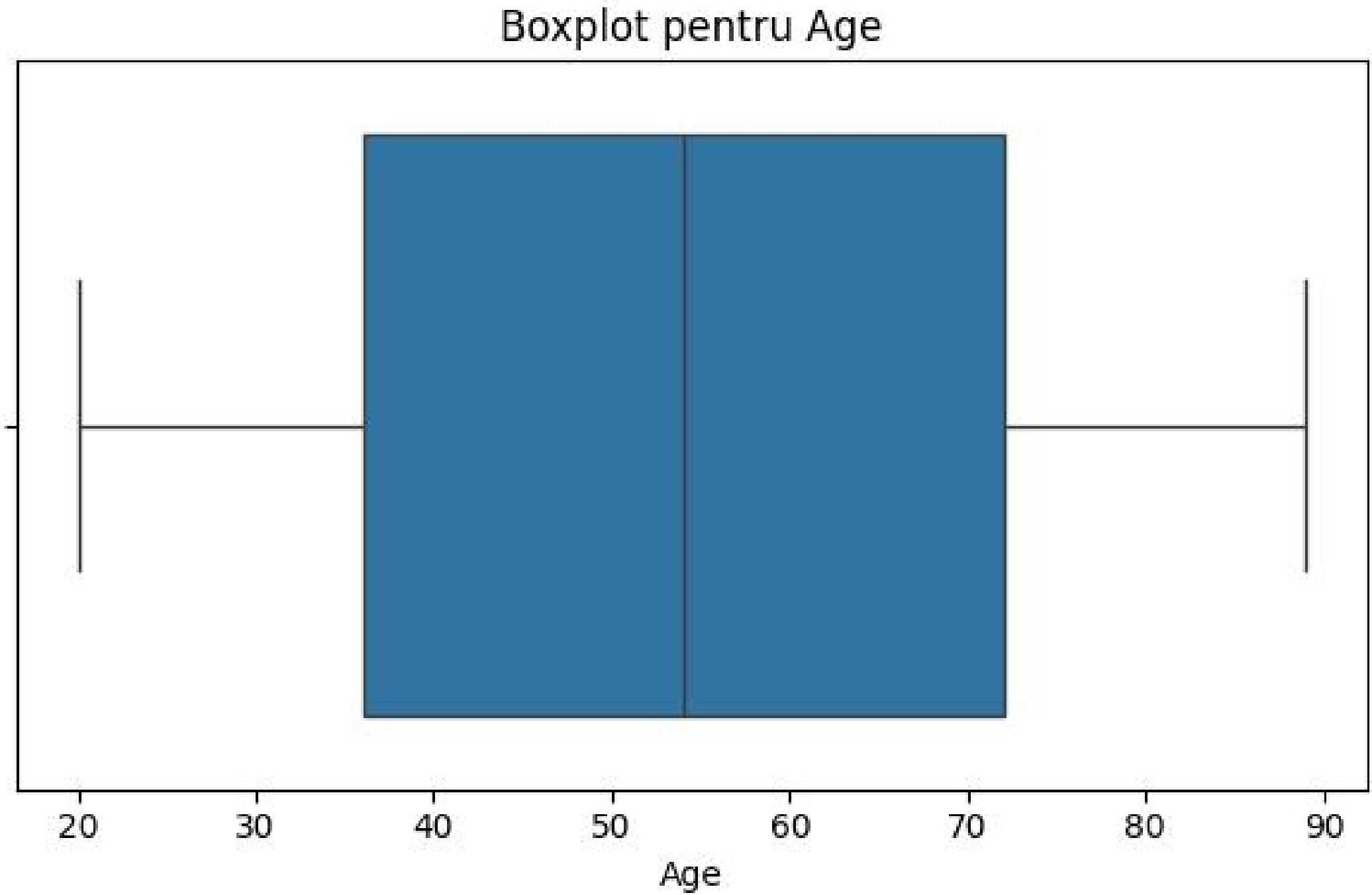
Una peste alta, din acest proiect am invatat atat cum sa gestionez un set de date, cat si cum sa-l interpretez si

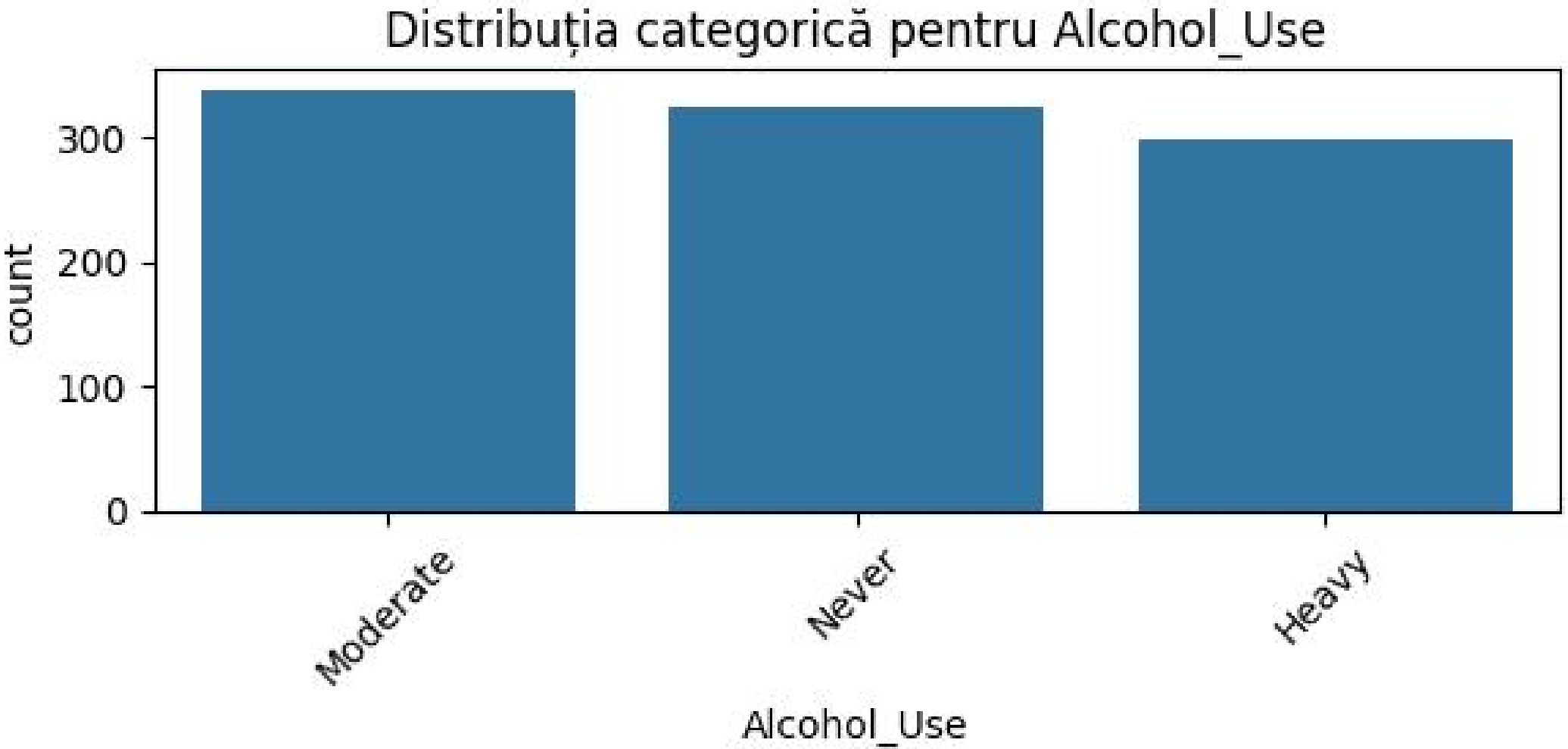
unde sa ma uit atunci cand apar erori sau dezechilibre.

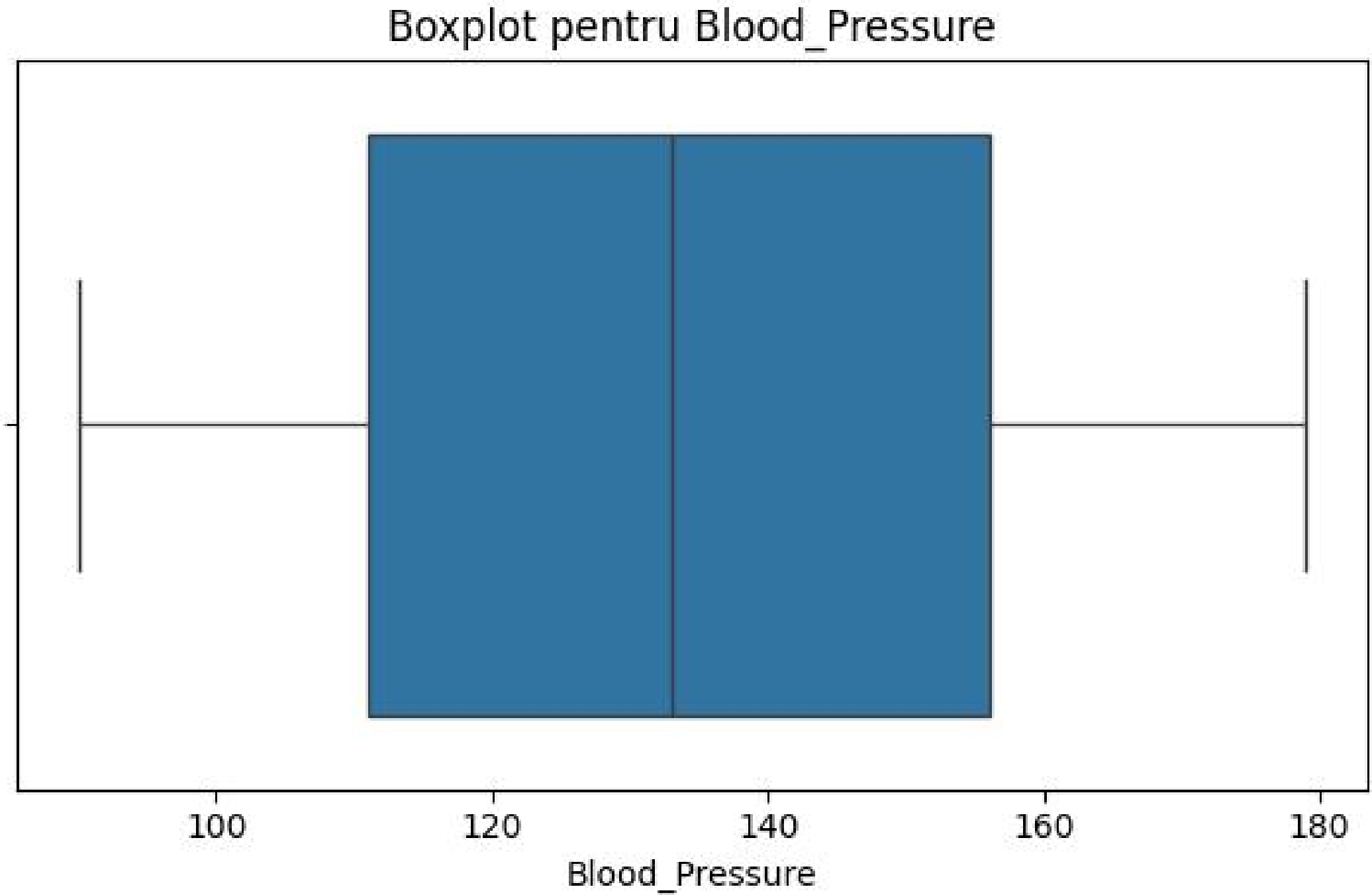
!!!Setul de date a fost generat de mine, ca experiment, impreuna cu Deepseek(de acolo si discorantele).

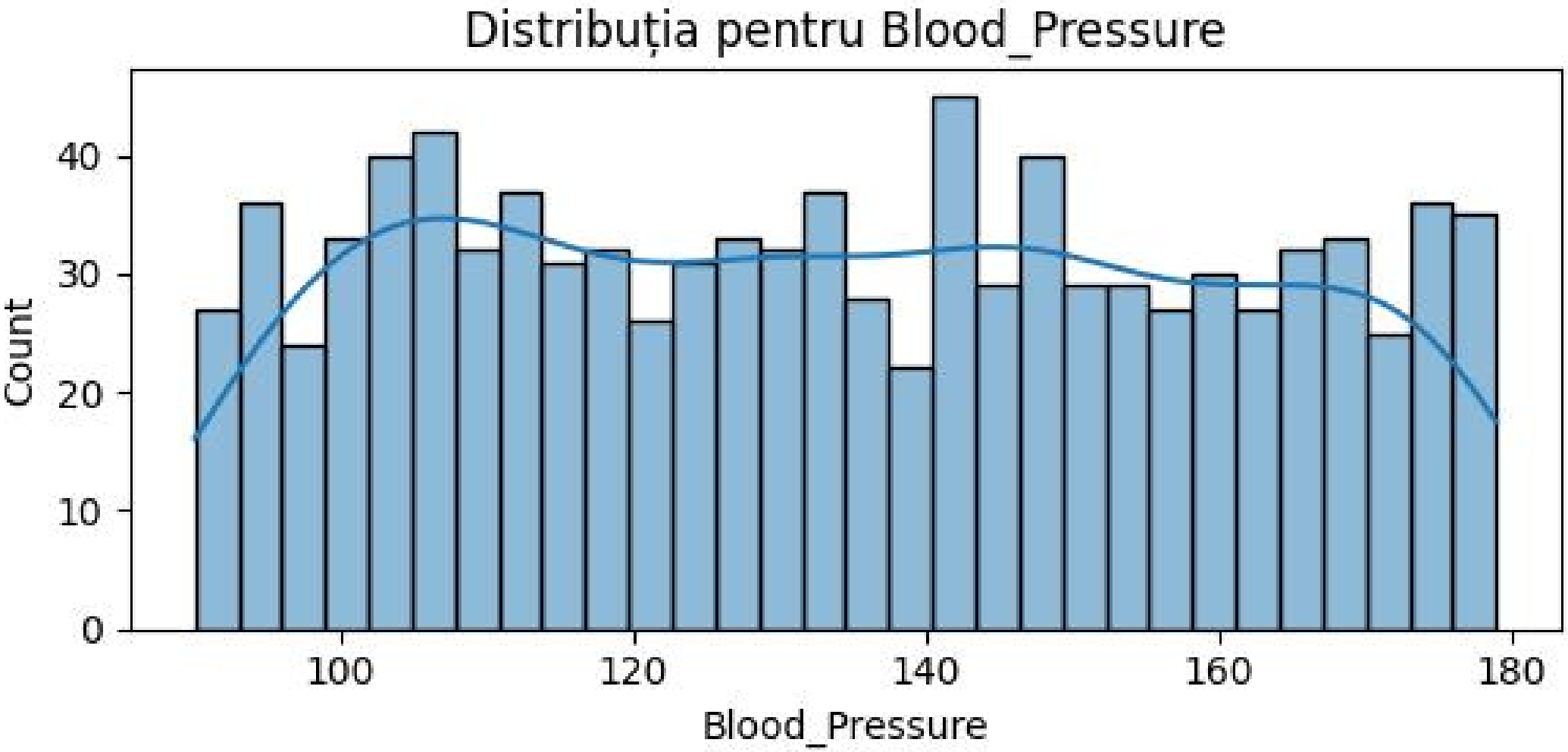
!!!Puteti gasi solutia mea si pe Github, aici: https://github.com/SanduStefan/PCLP3-Proiect.git

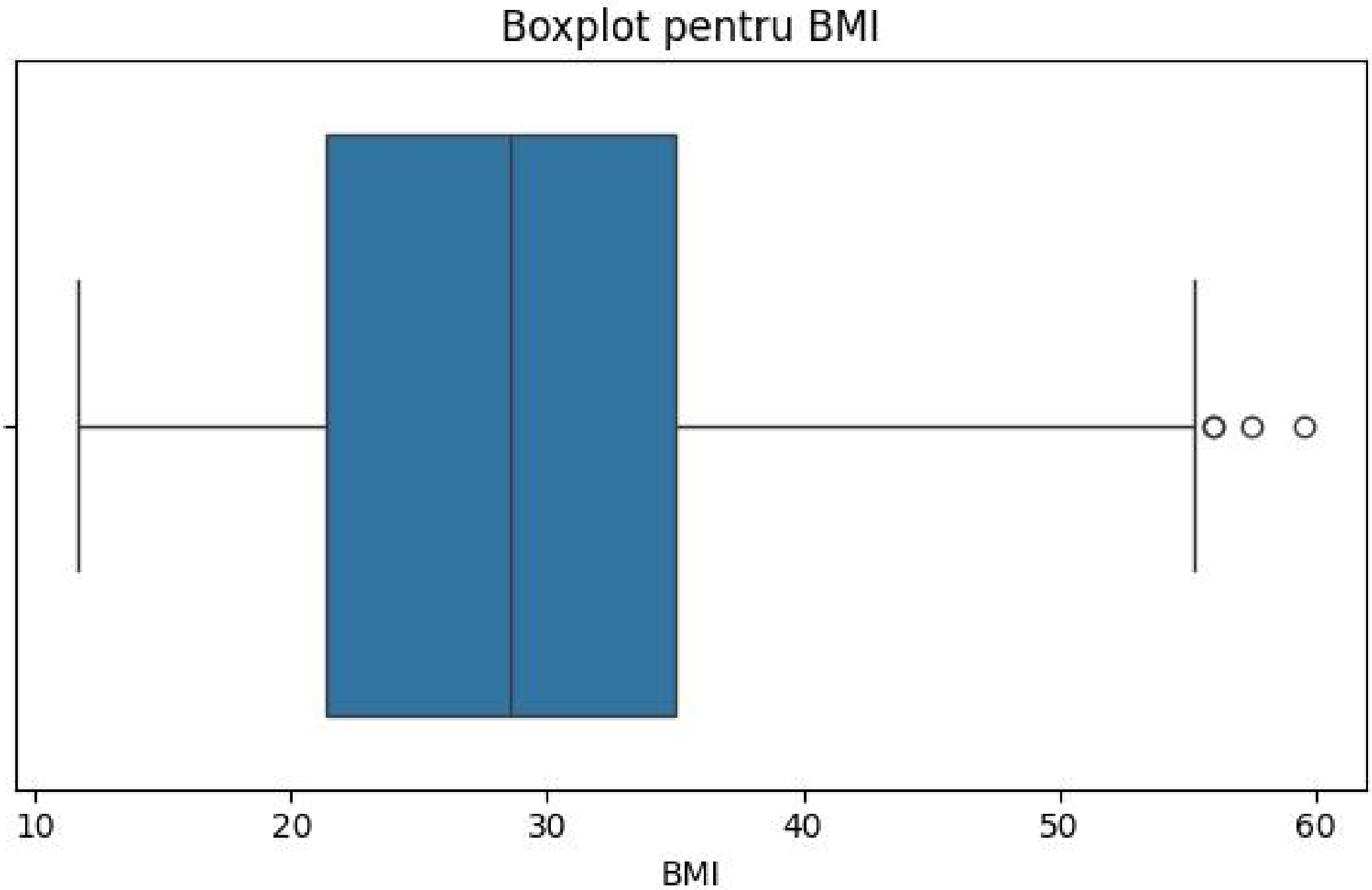

Distribuția pentru Age
Count
0
10
20
30
40
50
20
30
40
50
60
70
80
90
Age

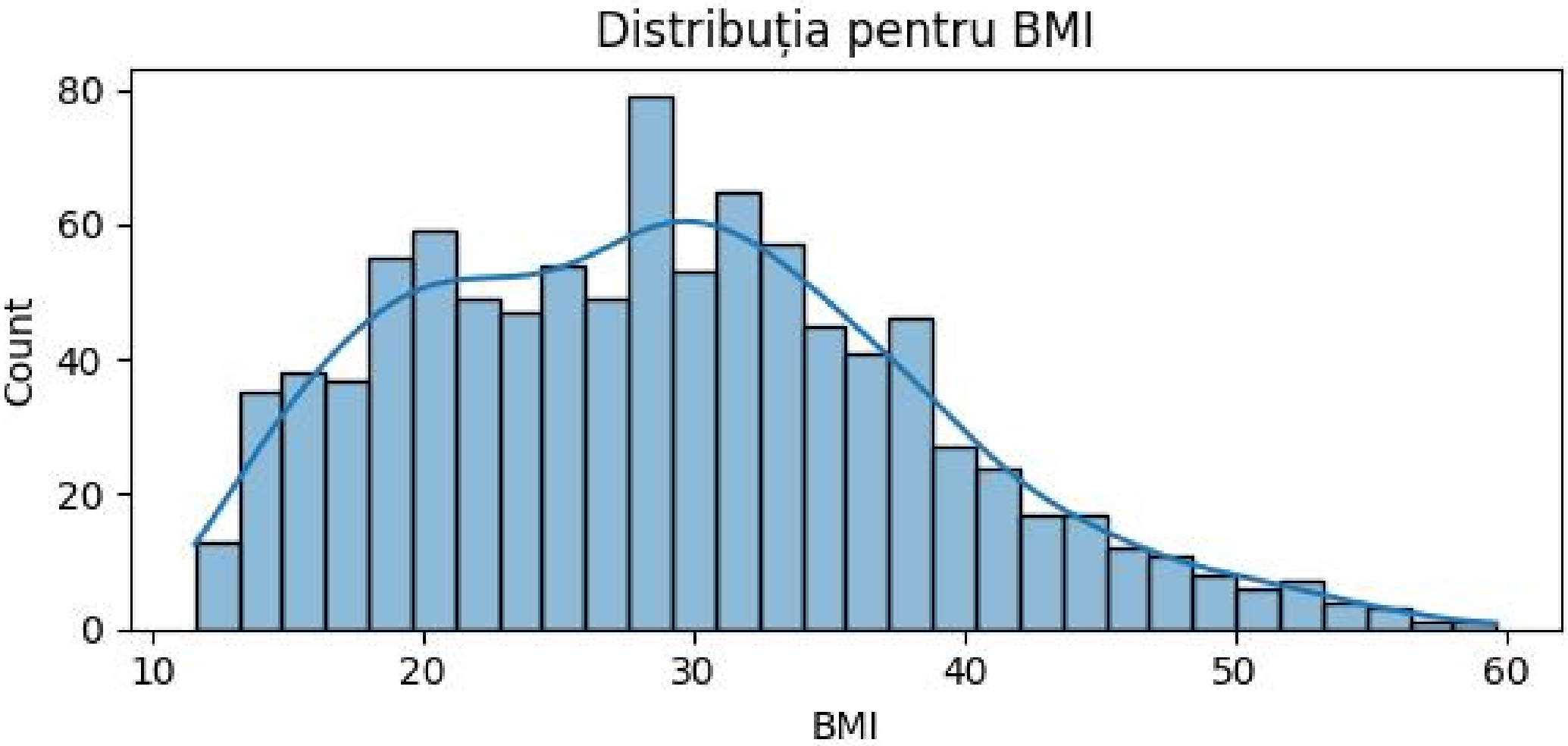

Boxplot pentru Age
20
30
40
50
60
70
80
90
Age

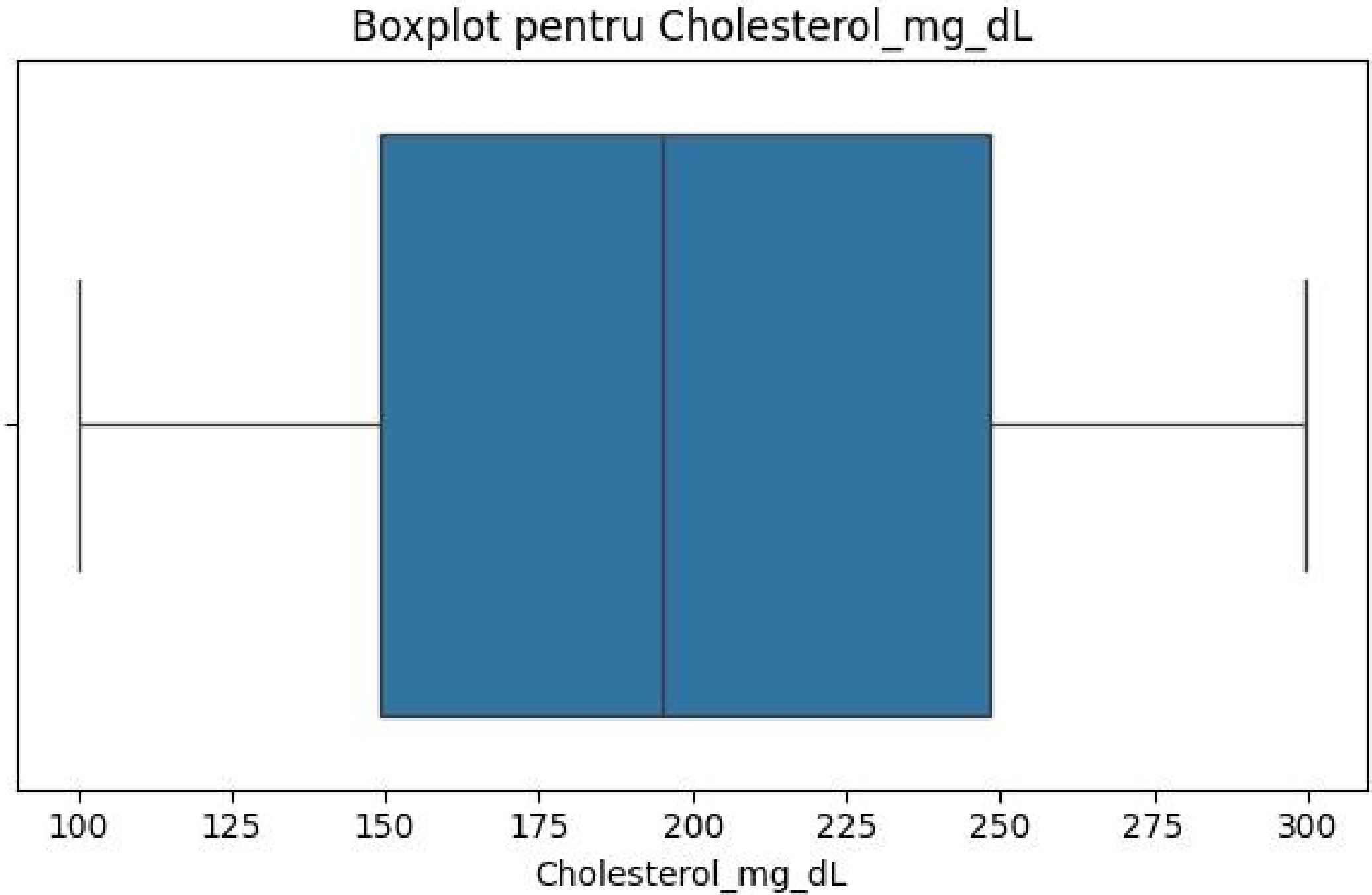

Distribuția categorică pentru Alcohol_Use
count
300
200
100
0
Moderate
Never
Heavy
Alcohol_Use

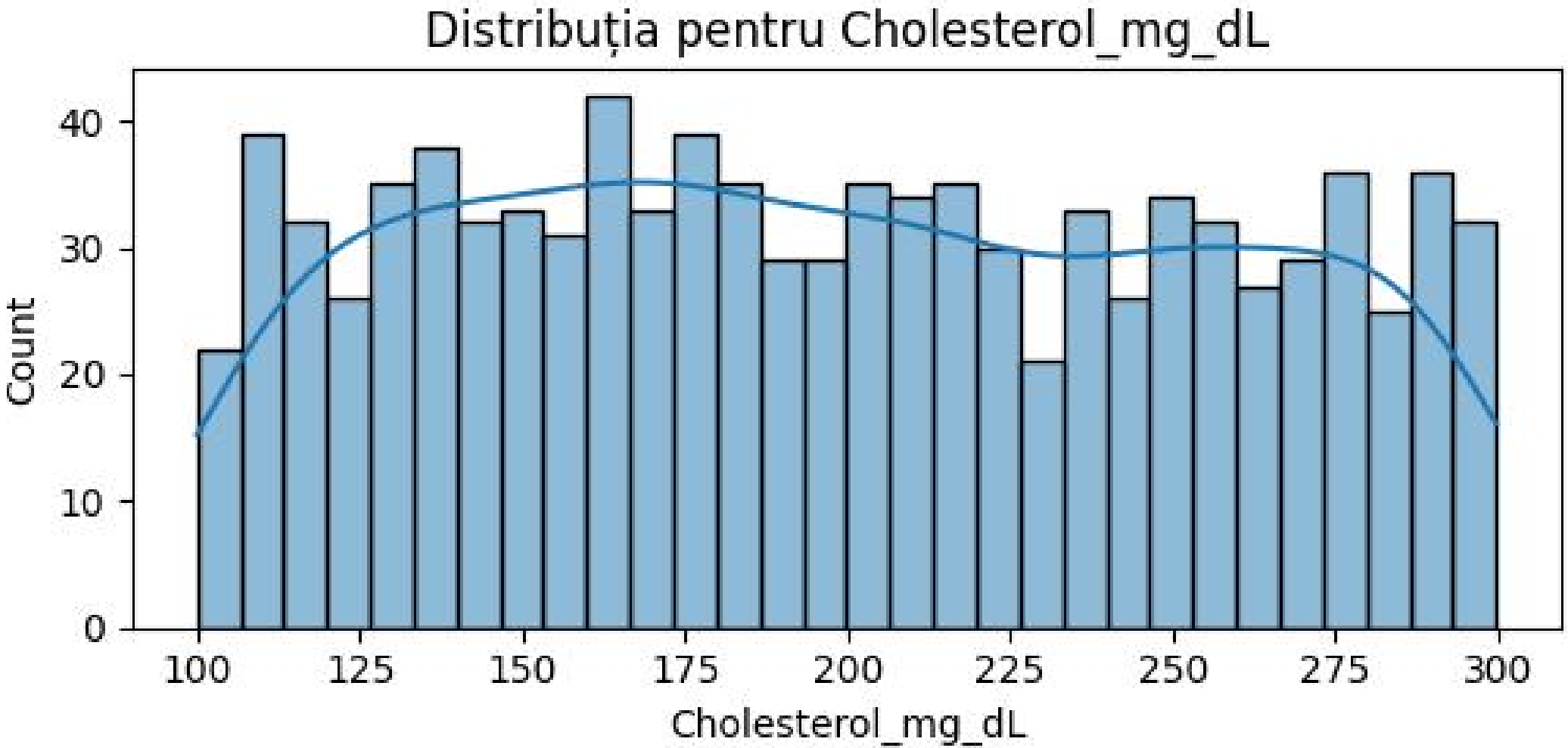

Boxplot pentru Blood_Pressure
100
120
140
160
180
Blood_Pressure

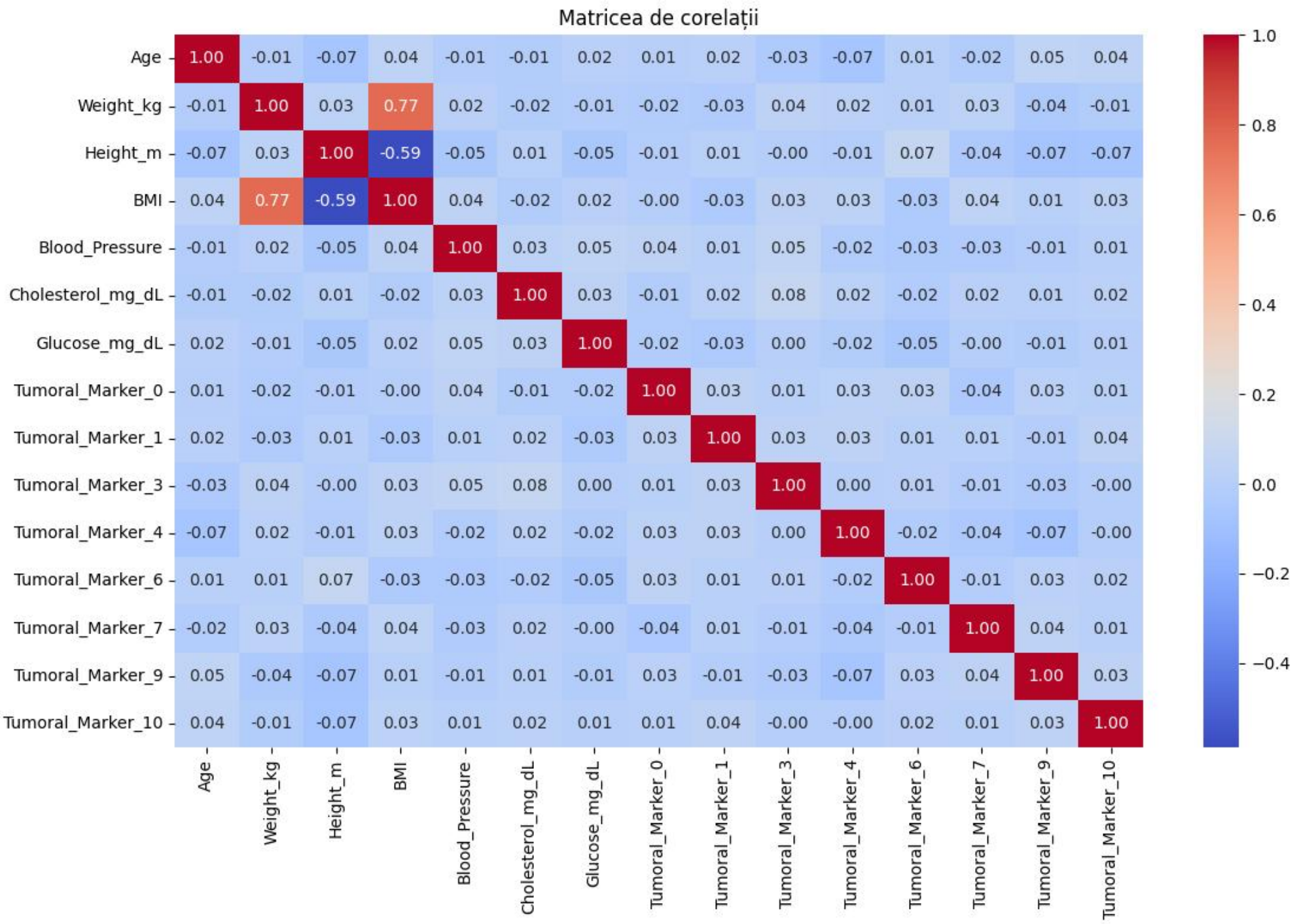

Distribuția pentru Blood_Pressure
Count
0
10
20
30
40
100
120
140
160
180
Blood_Pressure

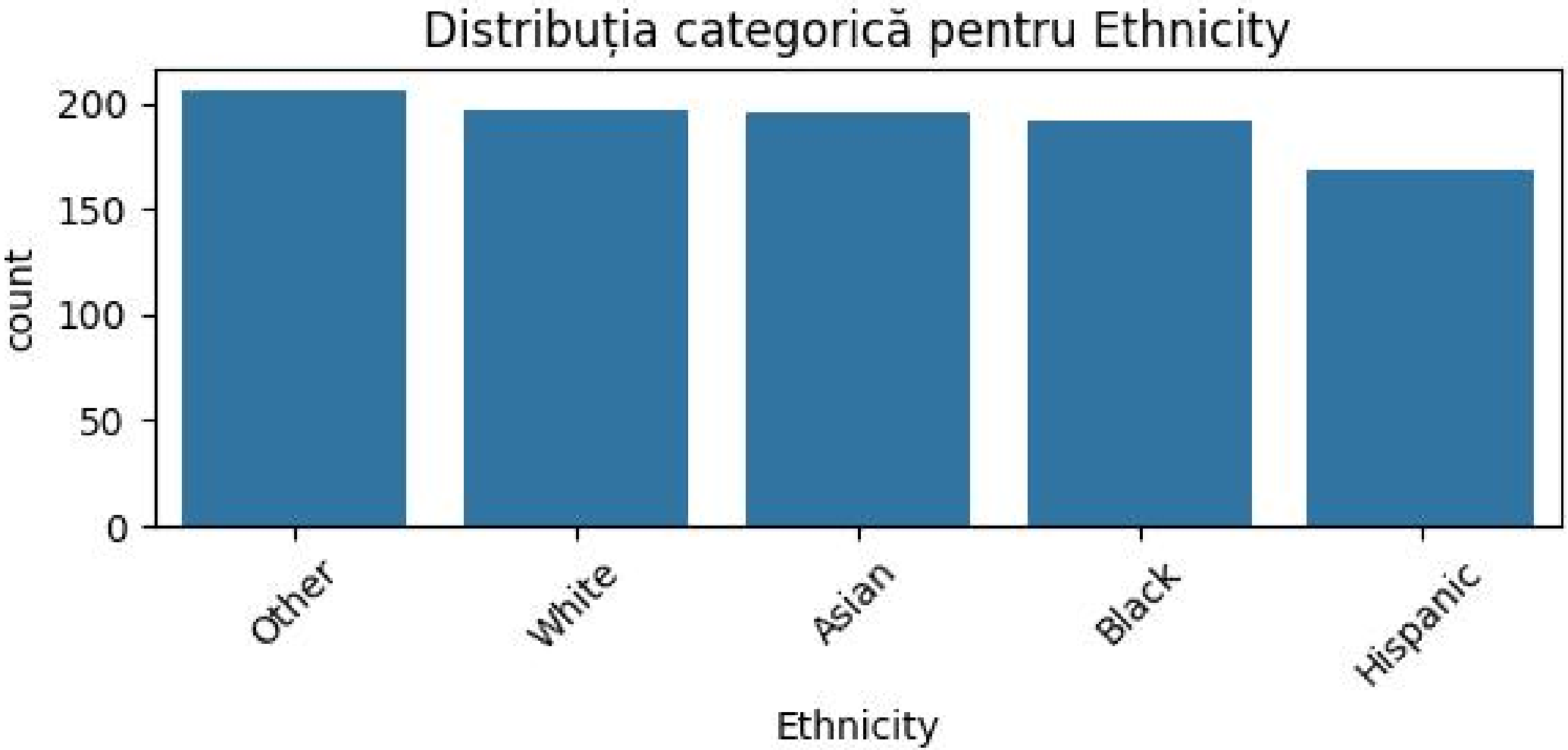

Boxplot pentru BMI
10
20
30
40
50
60
BMI

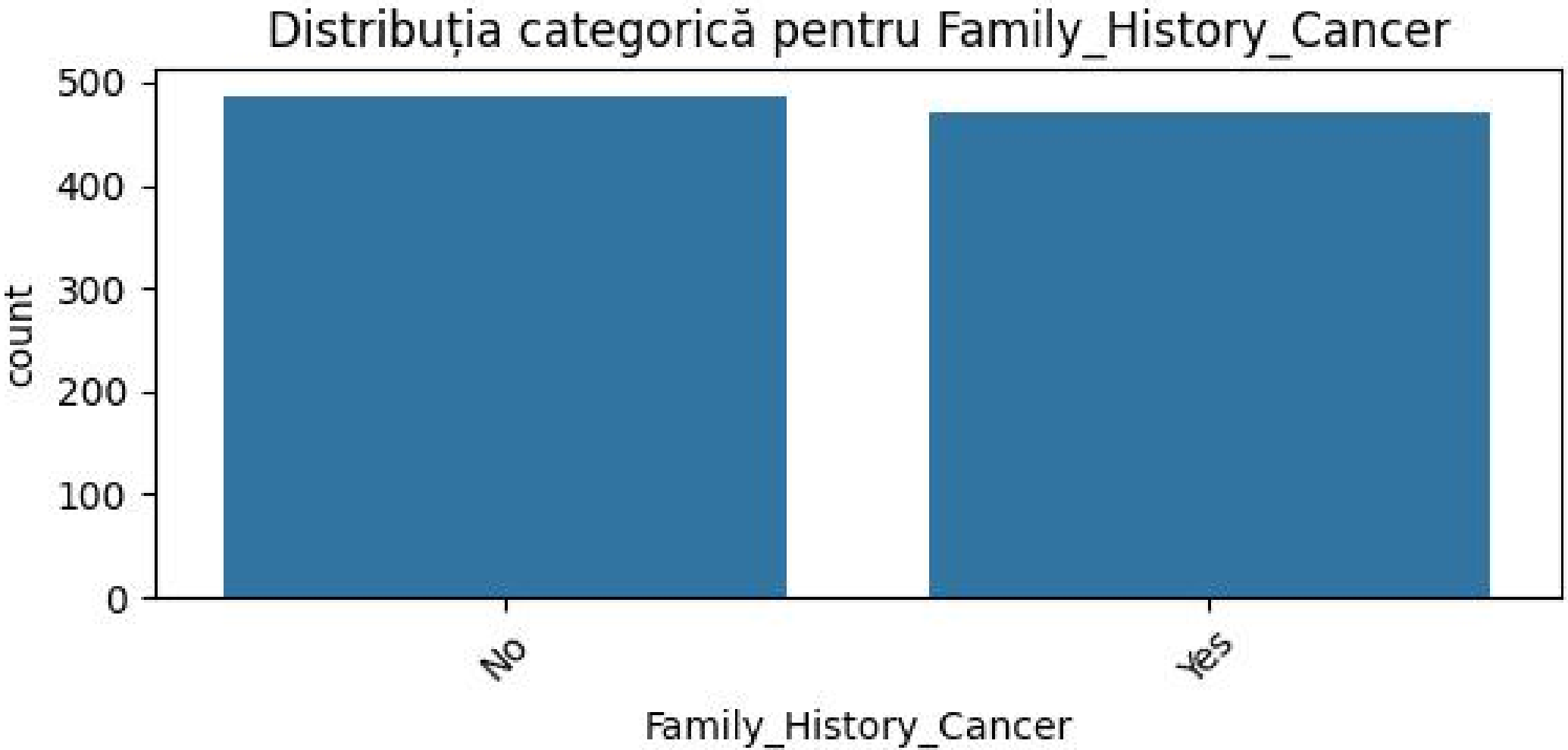

Distribuția pentru BMI
Count
0
20
40
60
80
10
20
30
40
50
60
BMI

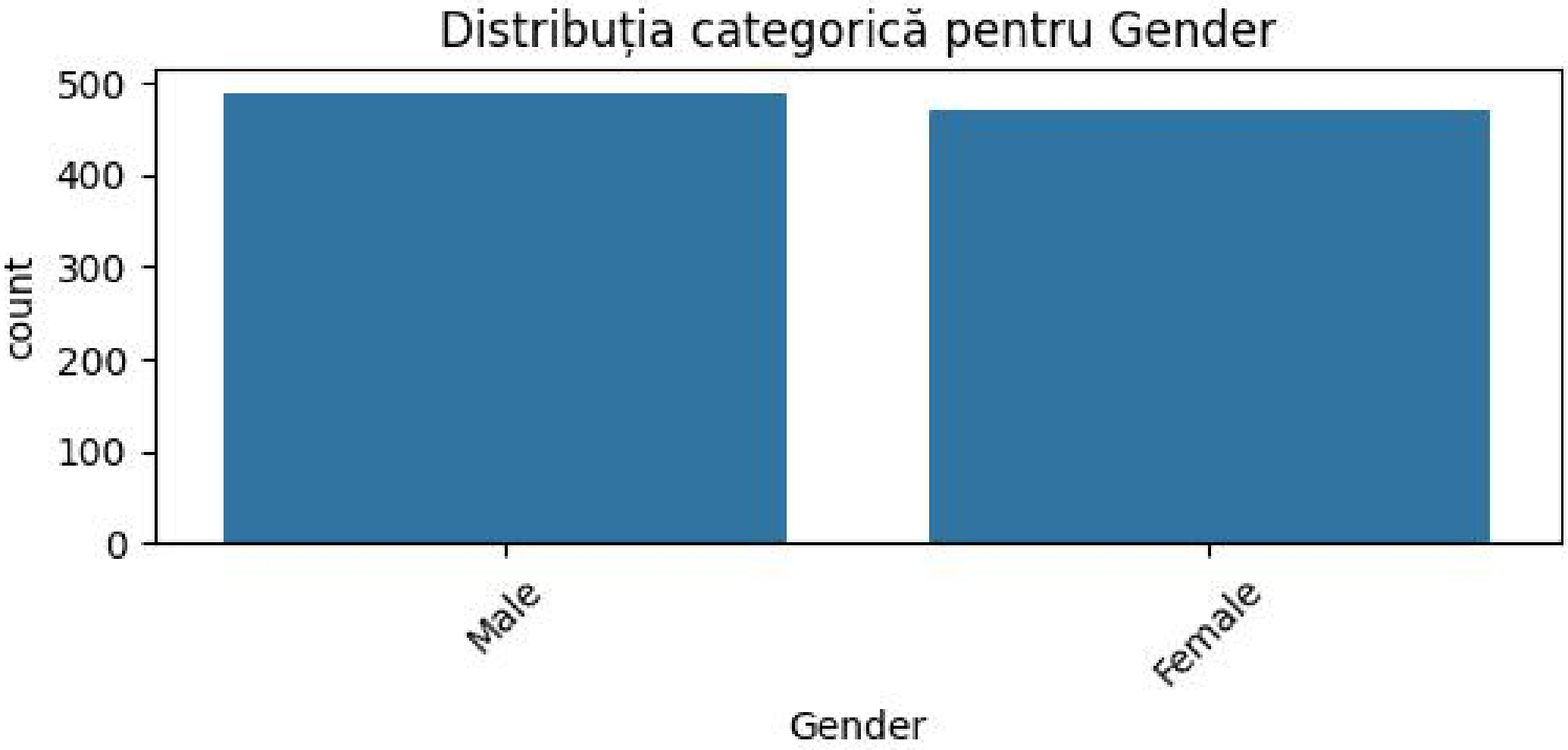

Boxplot pentru Cholesterol_mg_dL
100
125
150
175
200
225
250
275
300
Cholesterol_mg_dL

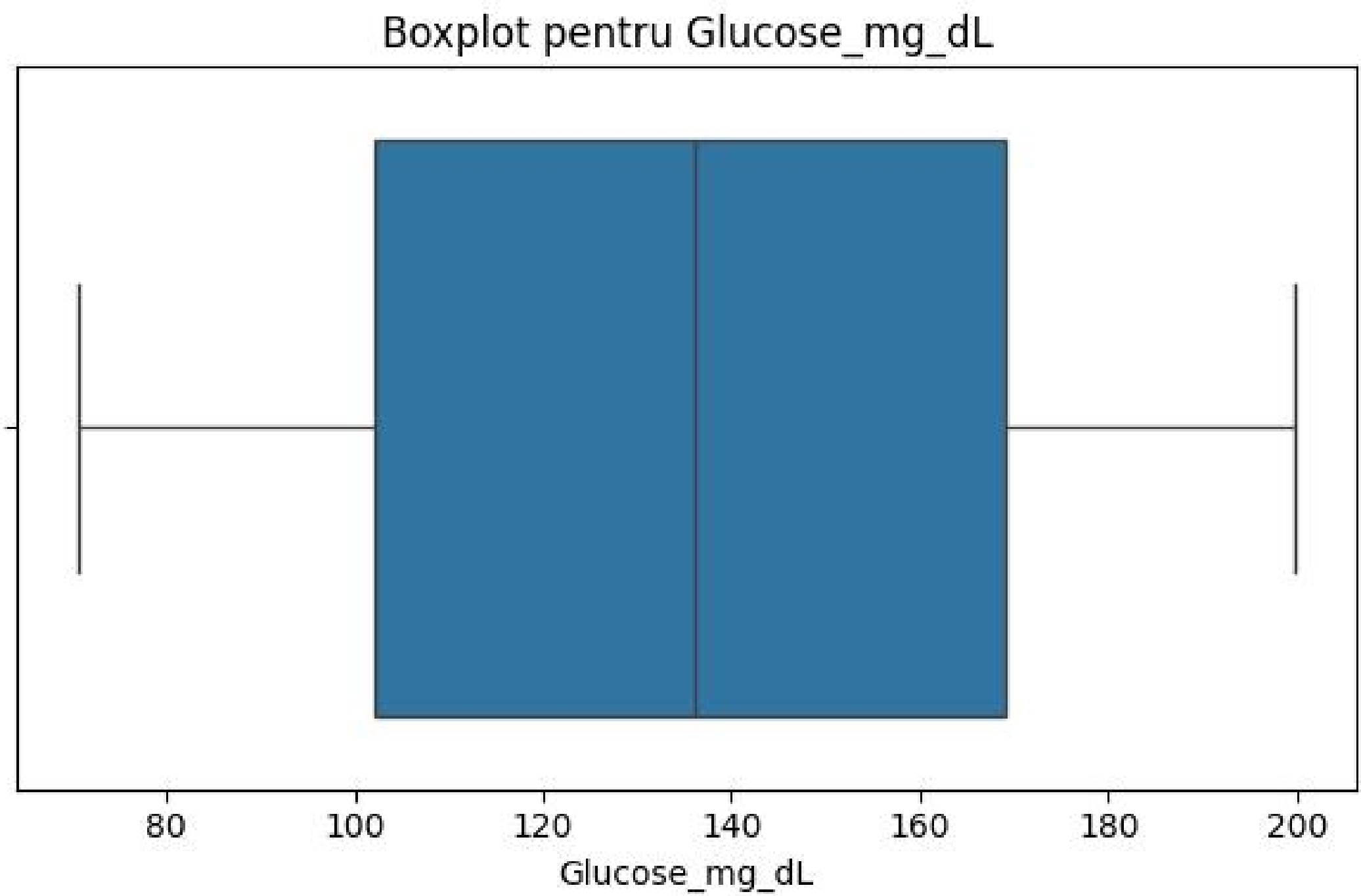

Distribuția pentru Cholesterol_mg_dL
Count
0
10
20
30
40
100
125
150
175
200
225
250
275
300
Cholesterol_mg_dL

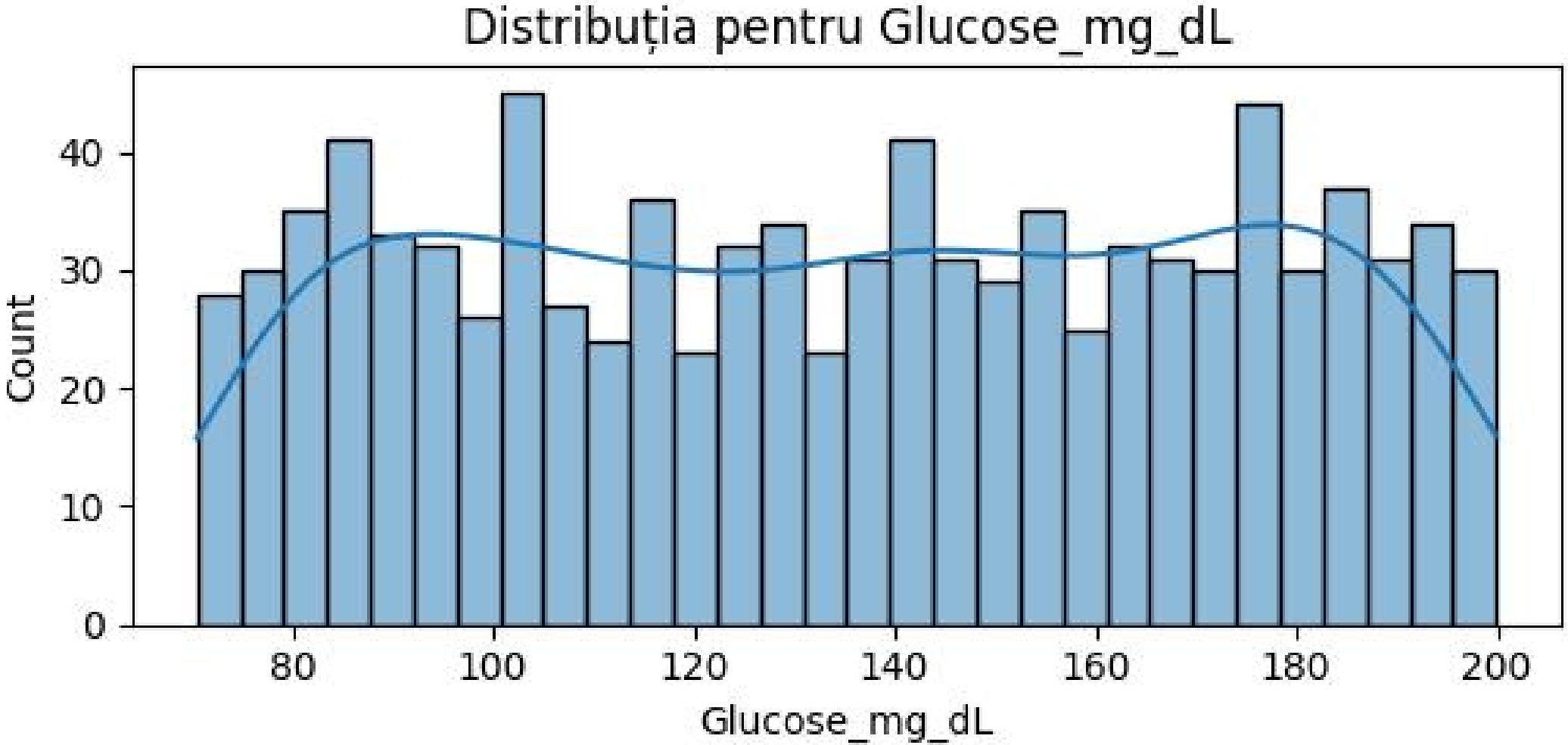

Matricea de corelații
Age 1.00 -0.01 -0.07 0.04 -0.01 -0.01 0.02 0.01 0.02 -0.03 -0.07 0.01 -0.02 0.05 0.04
Weight_kg -0.01 1.00 0.03 0.77 0.02 -0.02 -0.01 -0.02 -0.03 0.04 0.02 0.01 0.03 -0.04 -0.01
Height_m -0.07 0.03 1.00 -0.59 -0.05 0.01 -0.05 -0.01 0.01 -0.00 -0.01 0.07 -0.04 -0.07 -0.07
BMI 0.04 0.77 -0.59 1.00 0.04 -0.02 0.02 -0.00 -0.03 0.03 0.03 -0.03 0.04 0.01 0.03
Blood_Pressure -0.01 0.02 -0.05 0.04 1.00 0.03 0.05 0.04 0.01 0.05 -0.02 -0.03 -0.03 -0.01 0.01
Cholesterol_mg_dL -0.01 -0.02 0.01 -0.02 0.03 1.00 0.03 -0.01 0.02 0.08 0.02 -0.02 0.02 0.01 0.02
Glucose_mg_dL 0.02 -0.01 -0.05 0.02 0.05 0.03 1.00 -0.02 -0.03 0.00 -0.02 -0.05 -0.00 -0.01 0.01
Tumoral_Marker_0 0.01 -0.02 -0.01 -0.00 0.04 -0.01 -0.02 1.00 0.03 0.01 0.03 0.03 -0.04 0.03 0.01
Tumoral_Marker_1 0.02 -0.03 0.01 -0.03 0.01 0.02 -0.03 0.03 1.00 0.03 0.03 0.01 0.01 -0.01 0.04
Tumoral_Marker_3 -0.03 0.04 -0.00 0.03 0.05 0.08 0.00 0.01 0.03 1.00 0.00 0.01 -0.01 -0.03 -0.00
Tumoral_Marker_4 -0.07 0.02 -0.01 0.03 -0.02 0.02 -0.02 0.03 0.03 0.00 1.00 -0.02 -0.04 -0.07 -0.00
Tumoral_Marker_6 0.01 0.01 0.07 -0.03 -0.03 -0.02 -0.05 0.03 0.01 0.01 -0.02 1.00 -0.01 0.03 0.02
Tumoral_Marker_7 -0.02 0.03 -0.04 0.04 -0.03 0.02 -0.00 -0.04 0.01 -0.01 -0.04 -0.01 1.00 0.04 0.01
Tumoral_Marker_9 0.05 -0.04 -0.07 0.01 -0.01 0.01 -0.01 0.03 -0.01 -0.03 -0.07 0.03 0.04 1.00 0.03
Tumoral_Marker_10 0.04 -0.01 -0.07 0.03 0.01 0.02 0.01 0.01 0.04 -0.00 -0.00 0.02 0.01 0.03 1.00
Age
Weight_kg
Height_m
BMI
Blood_Pressure
Cholesterol_mg_dL
Glucose_mg_dL
Tumoral_Marker_0
Tumoral_Marker_1
Tumoral_Marker_3
Tumoral_Marker_4
Tumoral_Marker_6
Tumoral_Marker_7
Tumoral_Marker_9
Tumoral_Marker_10
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4

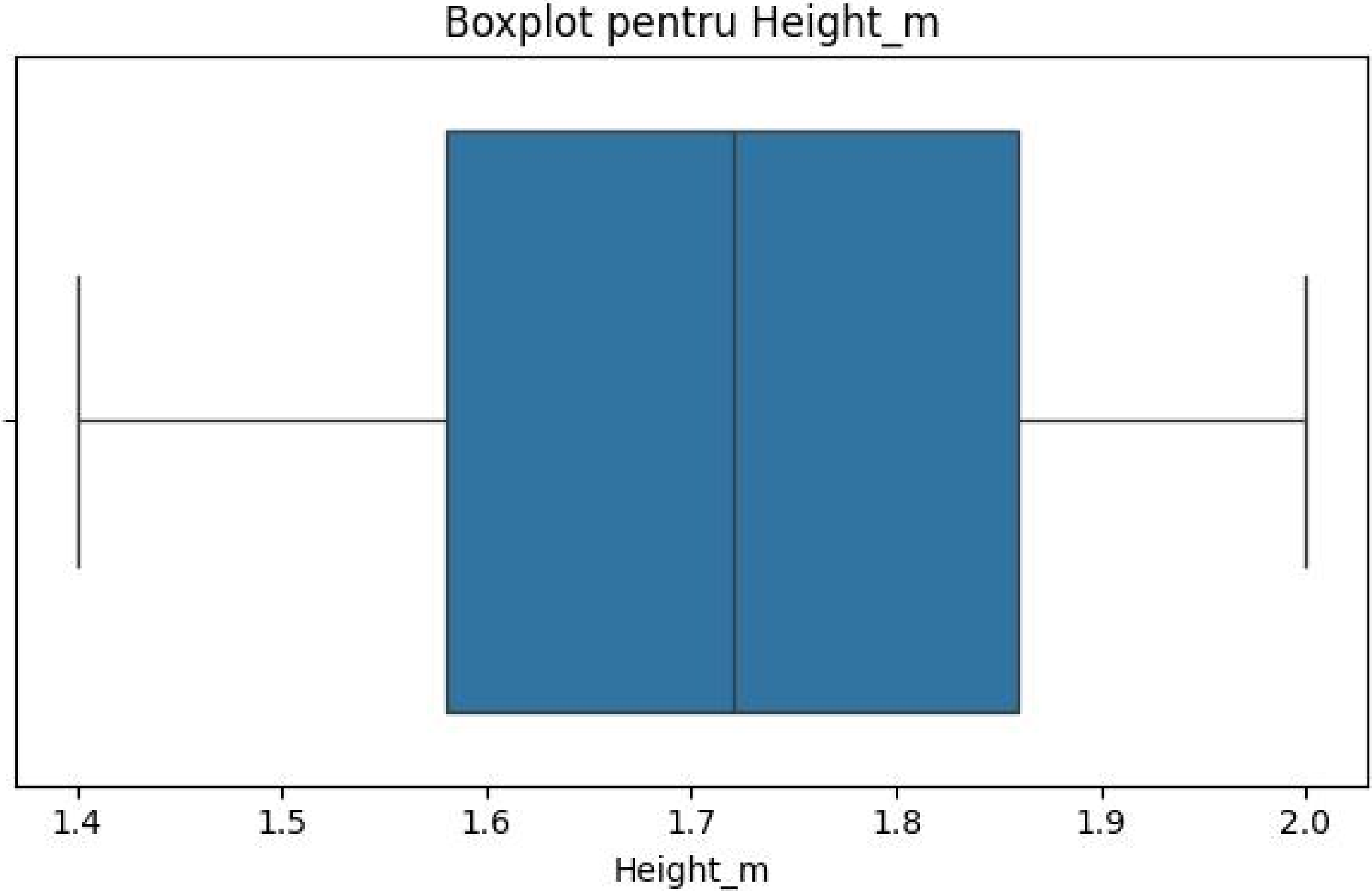

Distribuția categorică pentru Ethnicity
count
200
150
100
50
0
Other
White
Asian
Black
Hispanic
Ethnicity

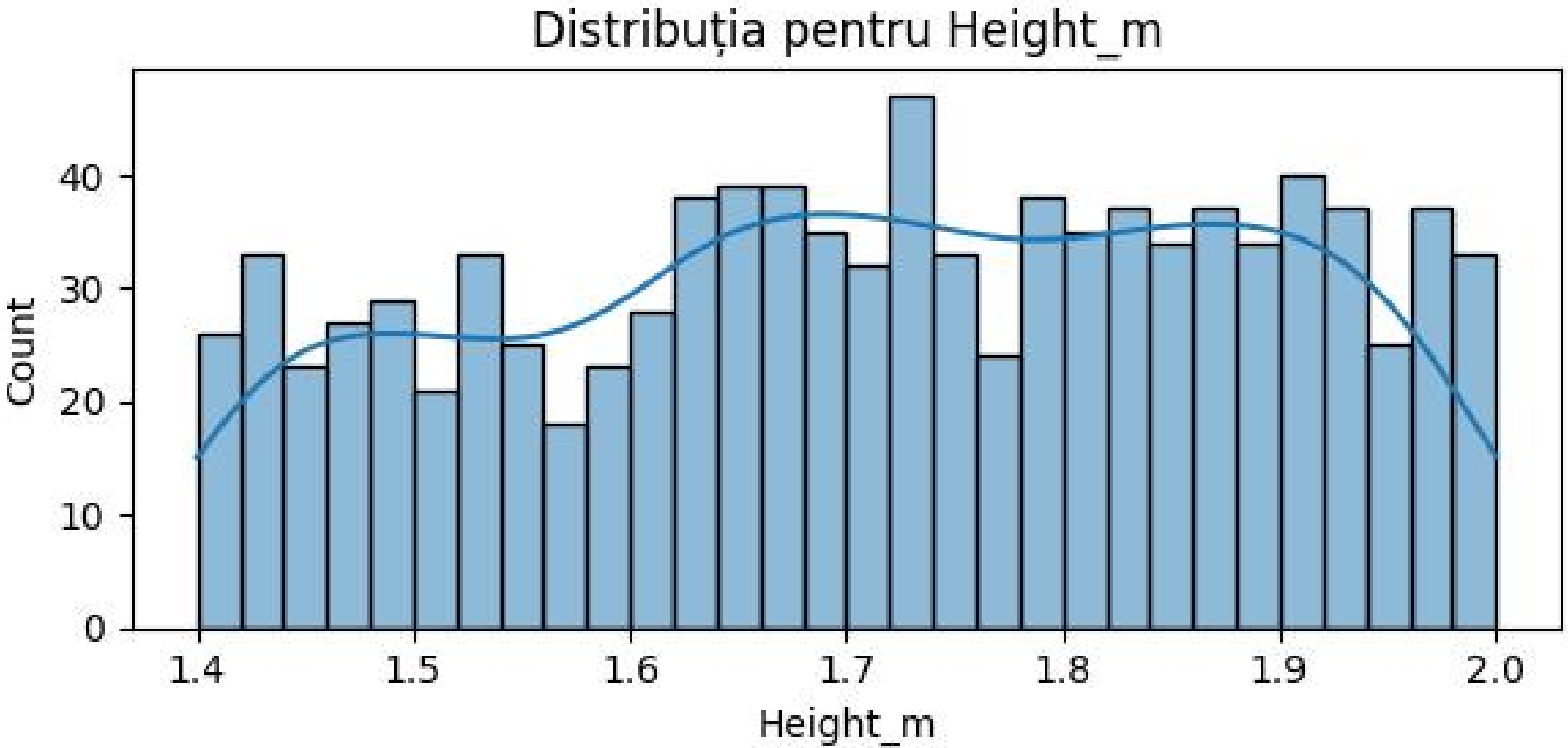

Distribuția categorică pentru Family_History_Cancer
500
400
300
200
100
0
count
No
Yes
Family_History_Cancer

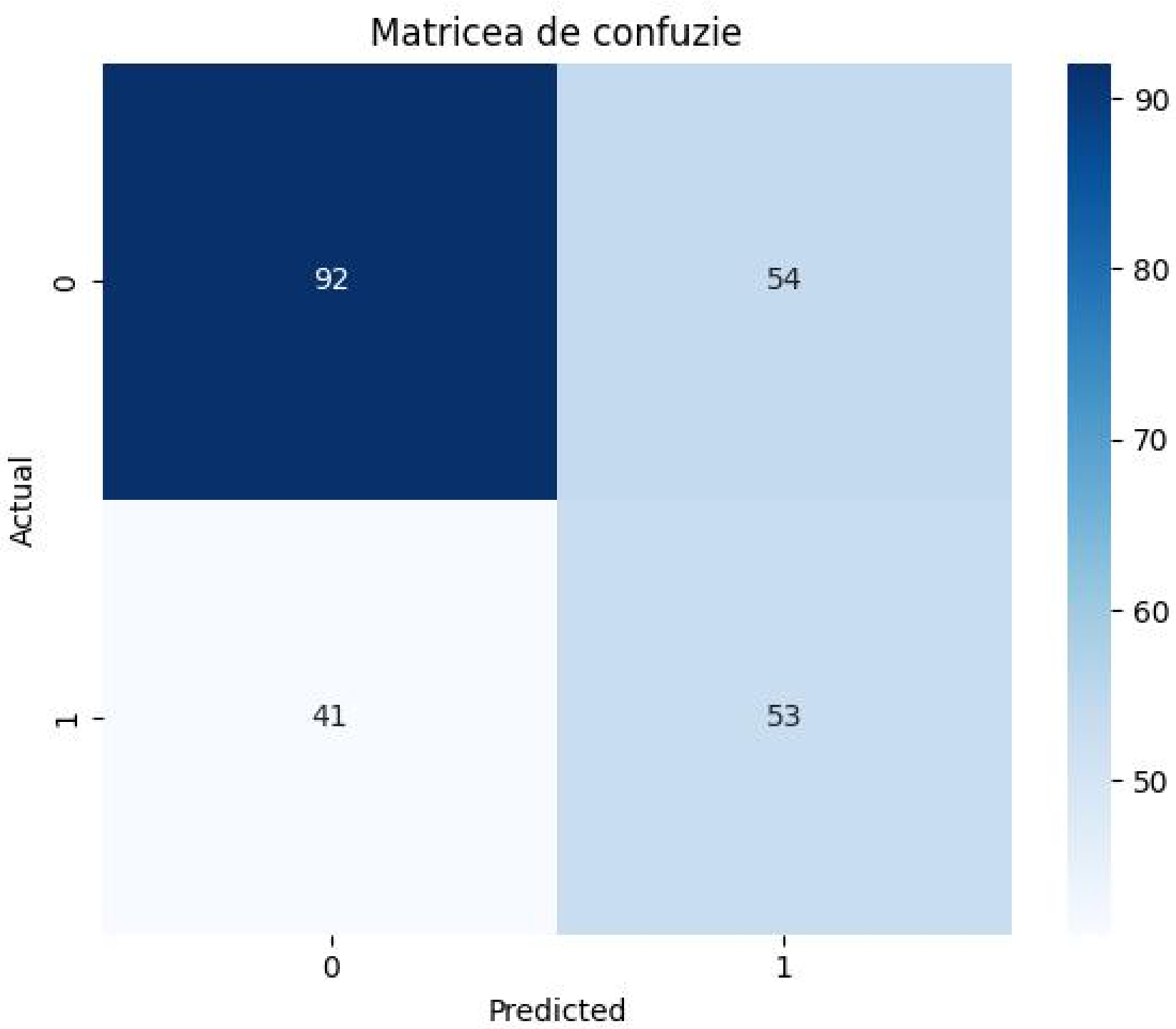

Distribuția categorică pentru Gender
count
500
400
300
200
100
0
Male
Female
Gender

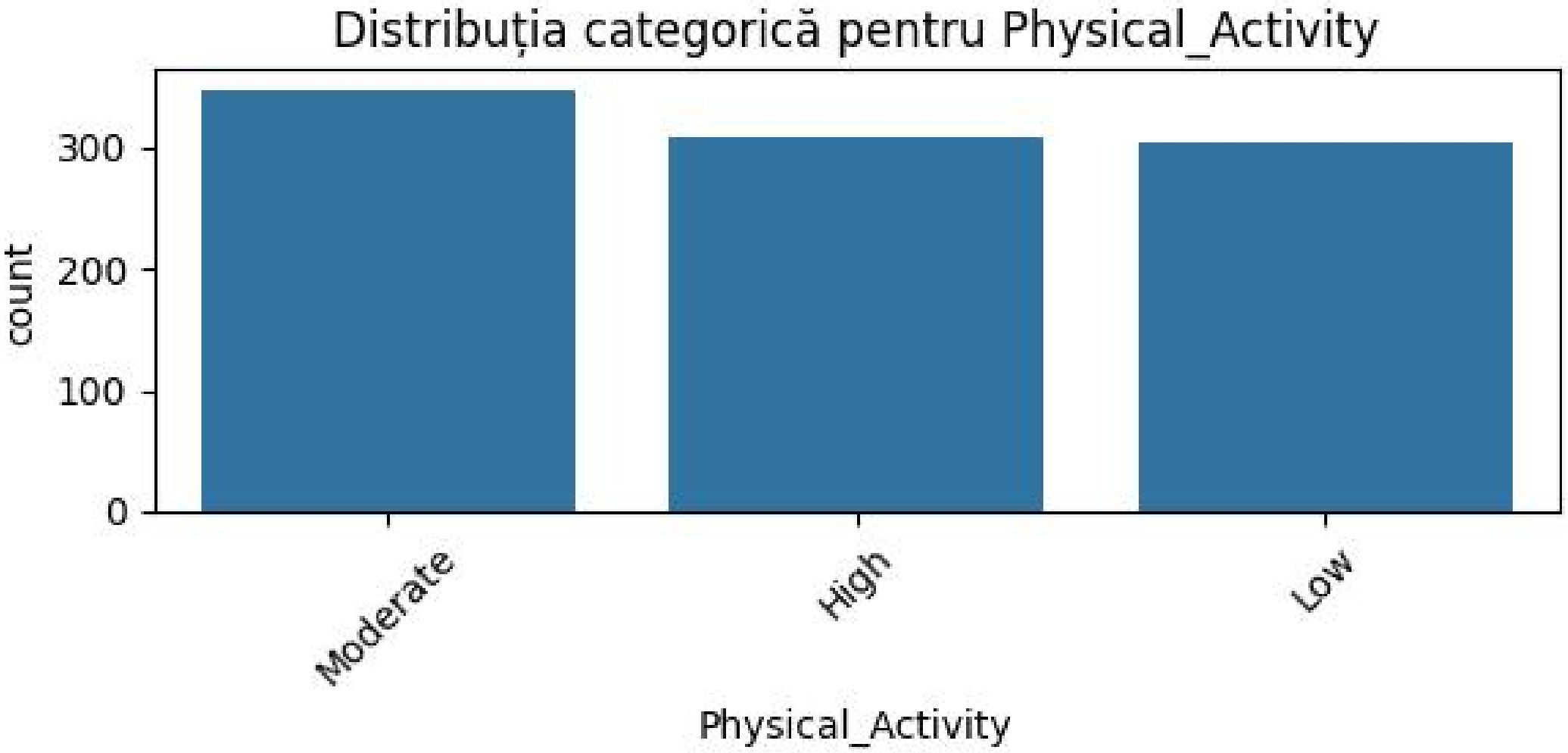

Boxplot pentru Glucose_mg_dL
80
100
120
140
160
180
200
Glucose_mg_dL

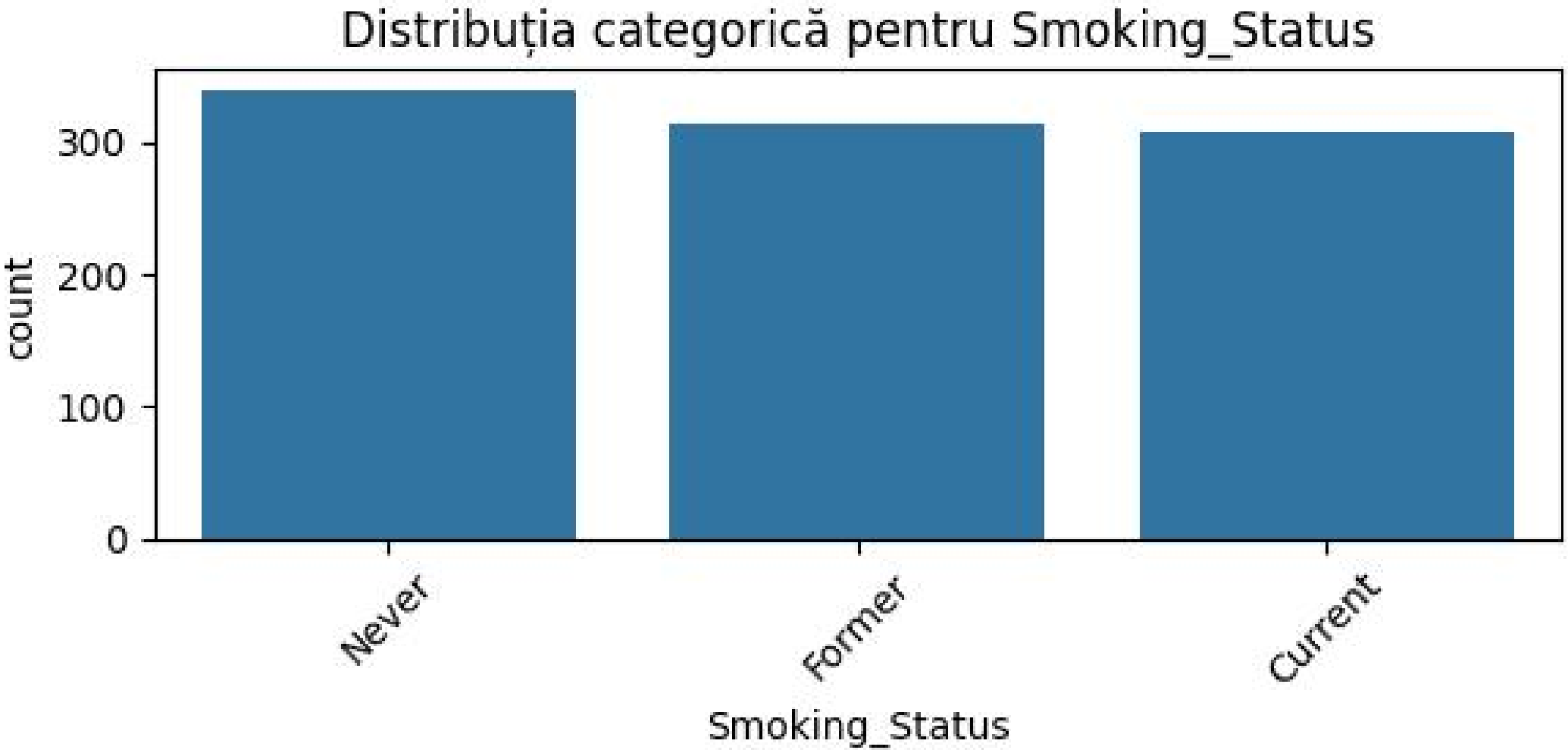

Distribuția pentru Glucose_mg_dL
Count
0
10
20
30
40
80
100
120
140
160
180
200
Glucose_mg_dL

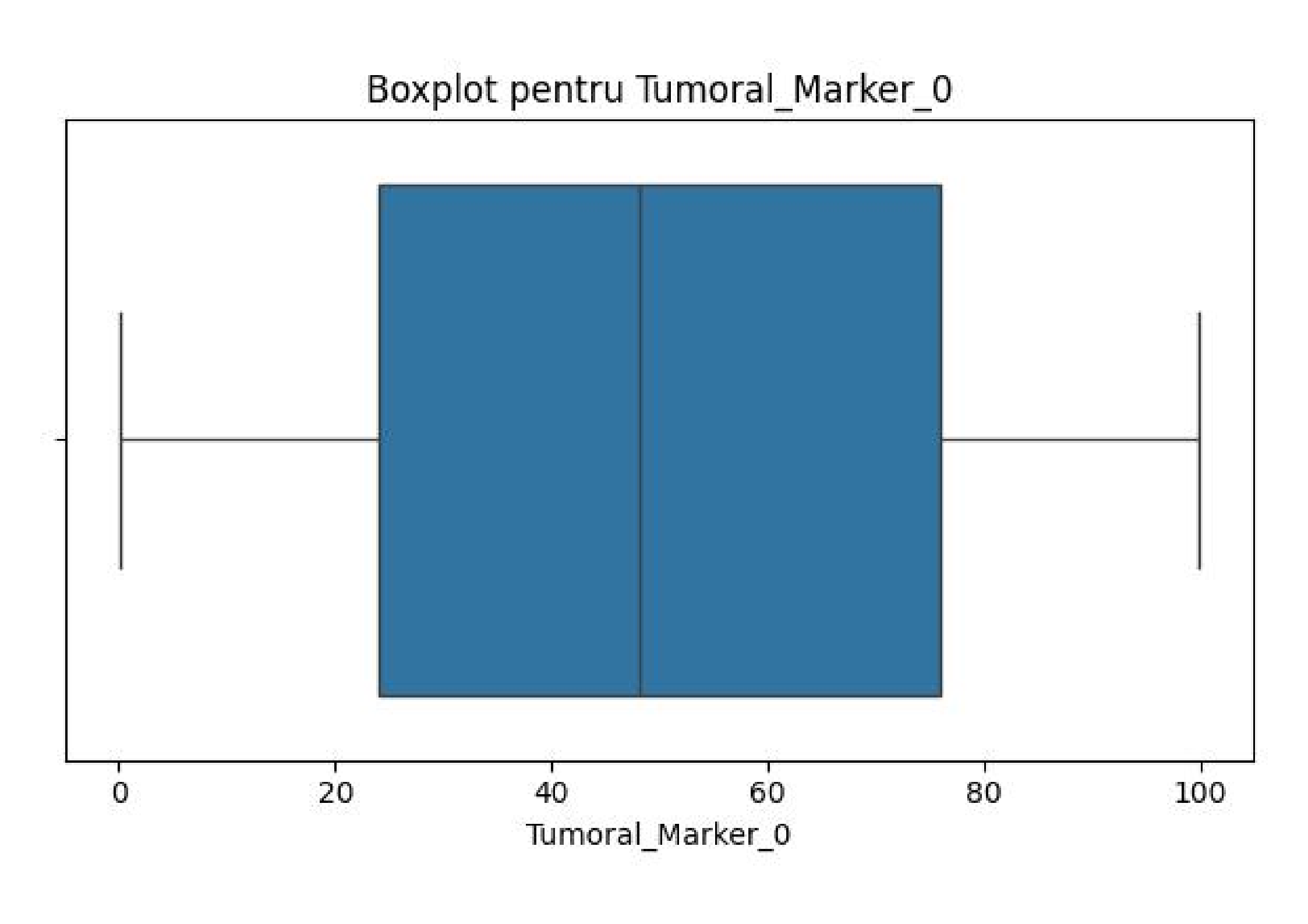

Boxplot pentru Height_m
1.4
1.5
1.6
1.7
1.8
1.9
2.0
Height_m

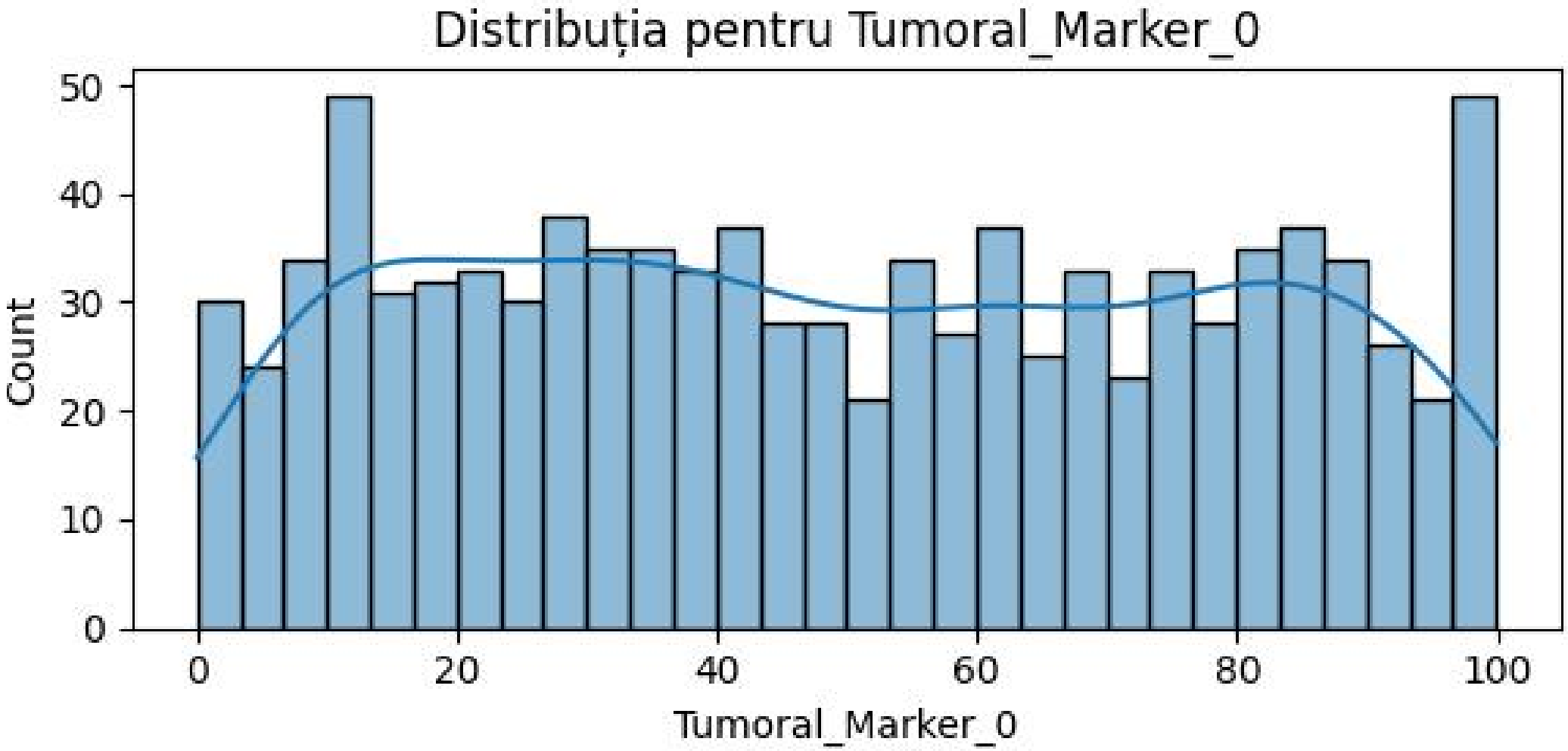

Distribuția pentru Height_m
Count
0
10
20
30
40
1.4
1.5
1.6
1.7
1.8
1.9
2.0
Height_m

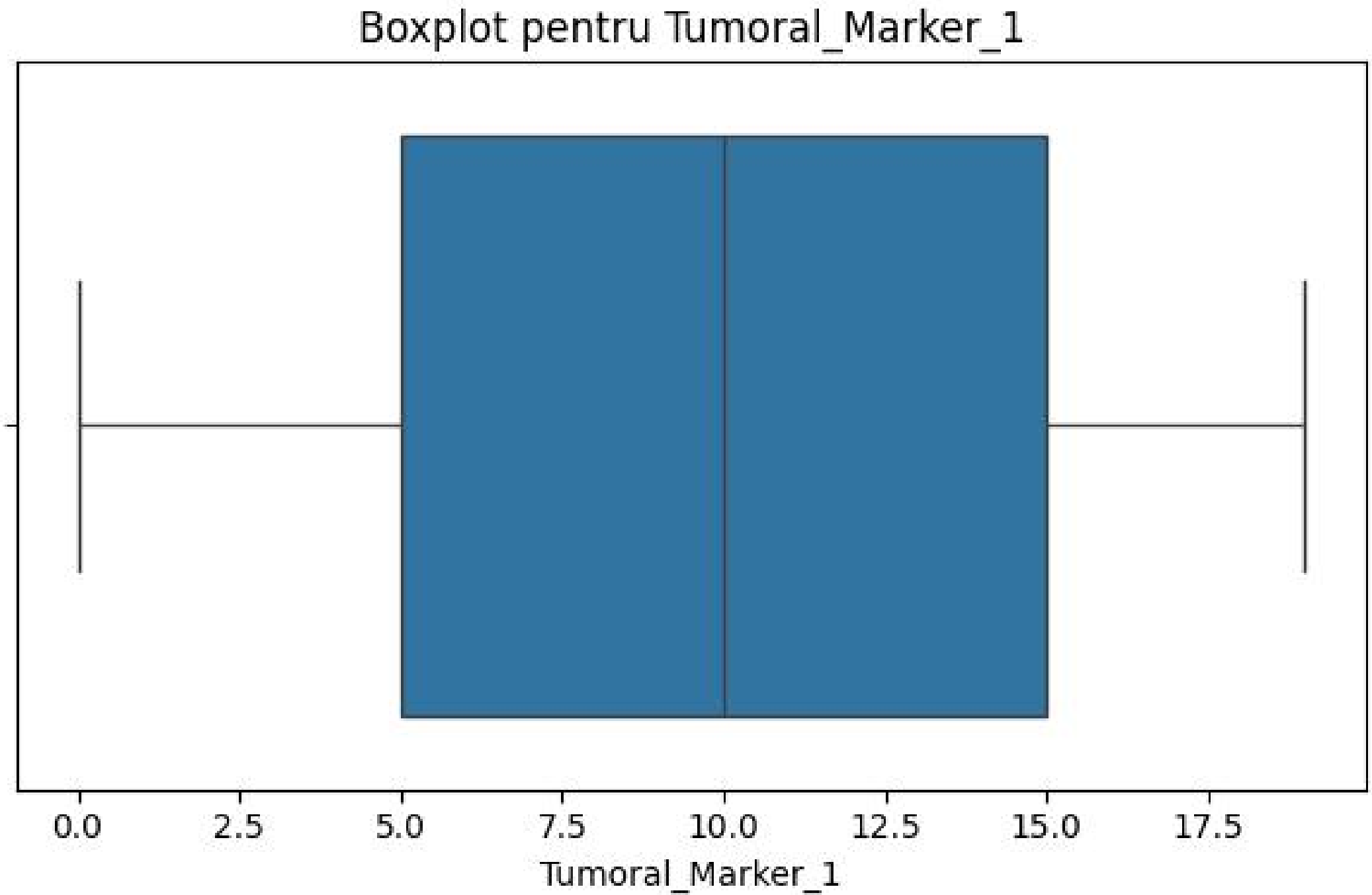

Matricea de confuzie
92
54
41
53
0
1
Actual
0
1
Predicted
90
80
70
60
50

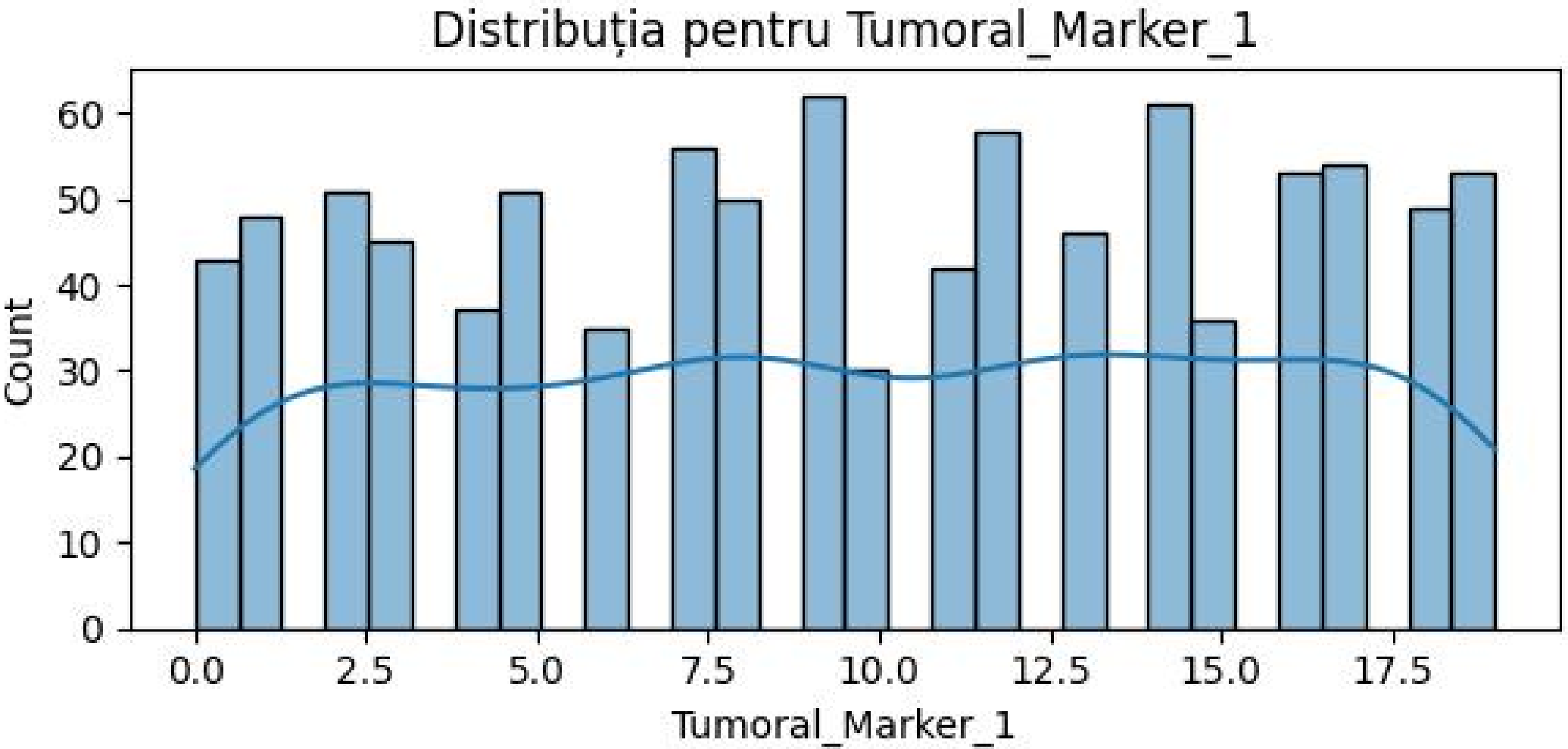

Distribuția categorică pentru Physical_Activity
count
300
200
100
0
Moderate
High
Low
Physical_Activity

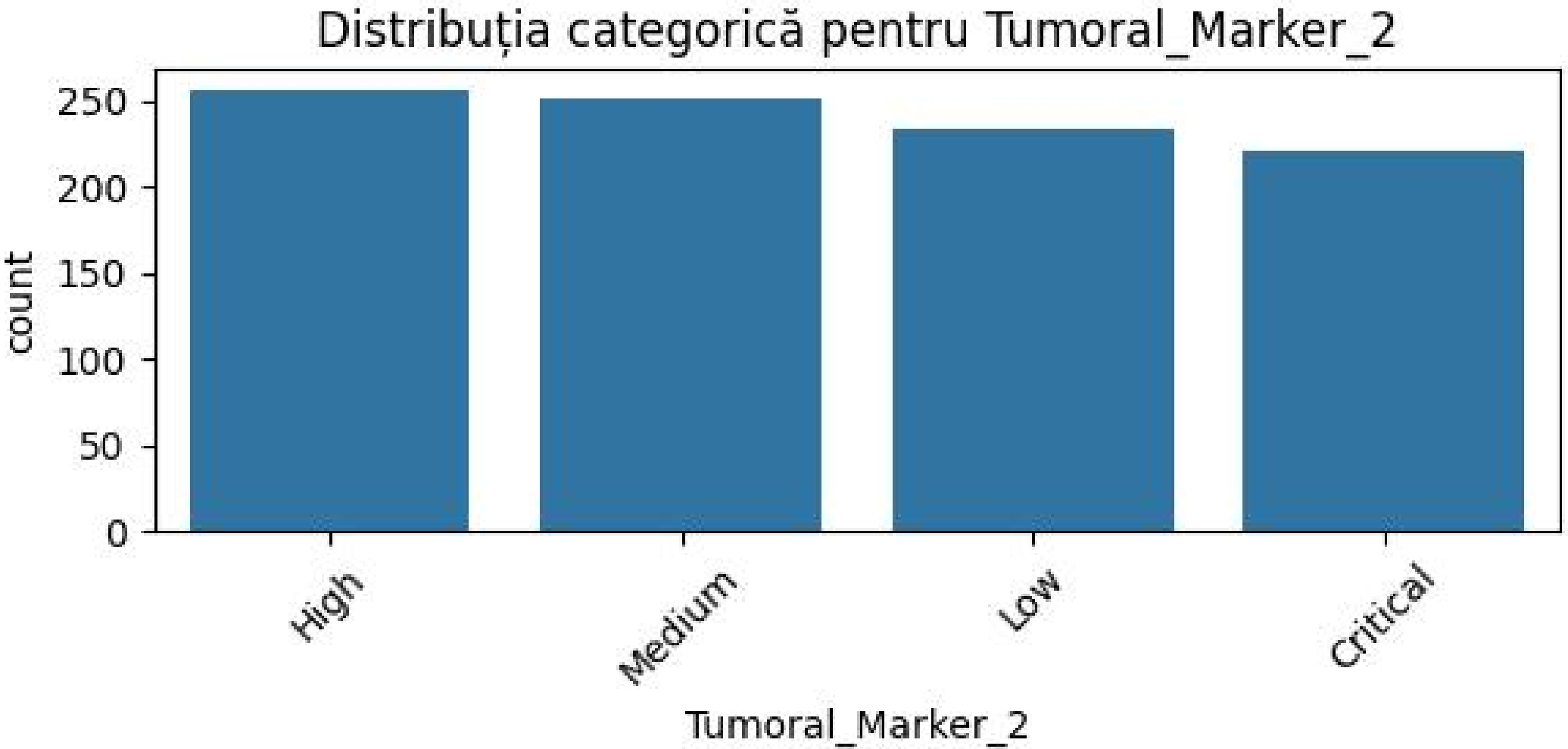

Distribuția categorică pentru Smoking_Status
count
300
200
100
0
Never
Former
Current
Smoking_Status

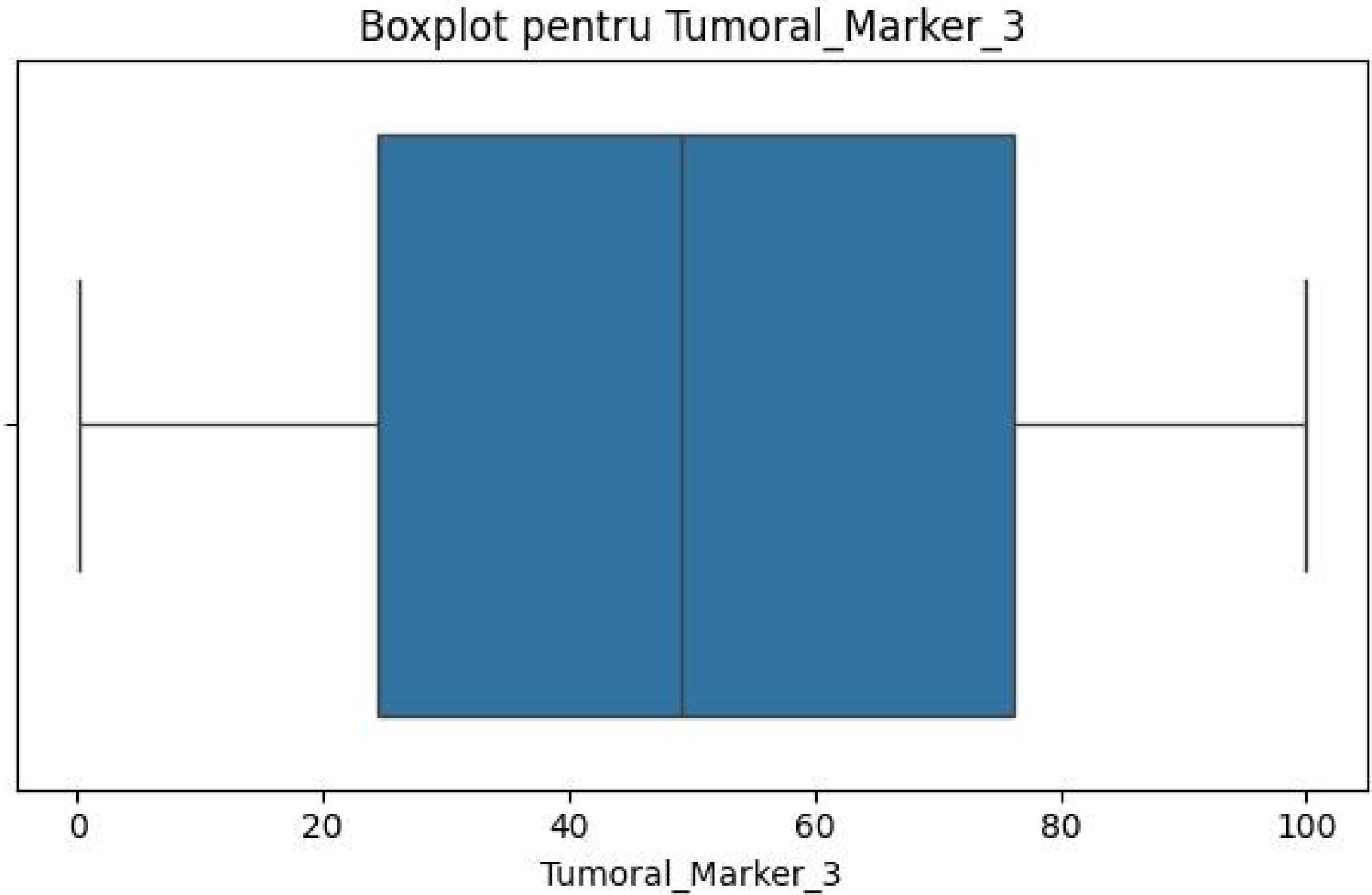

Boxplot pentru Tumoral_Marker_0
0
20
40
60
80
100
Tumoral_Marker_0

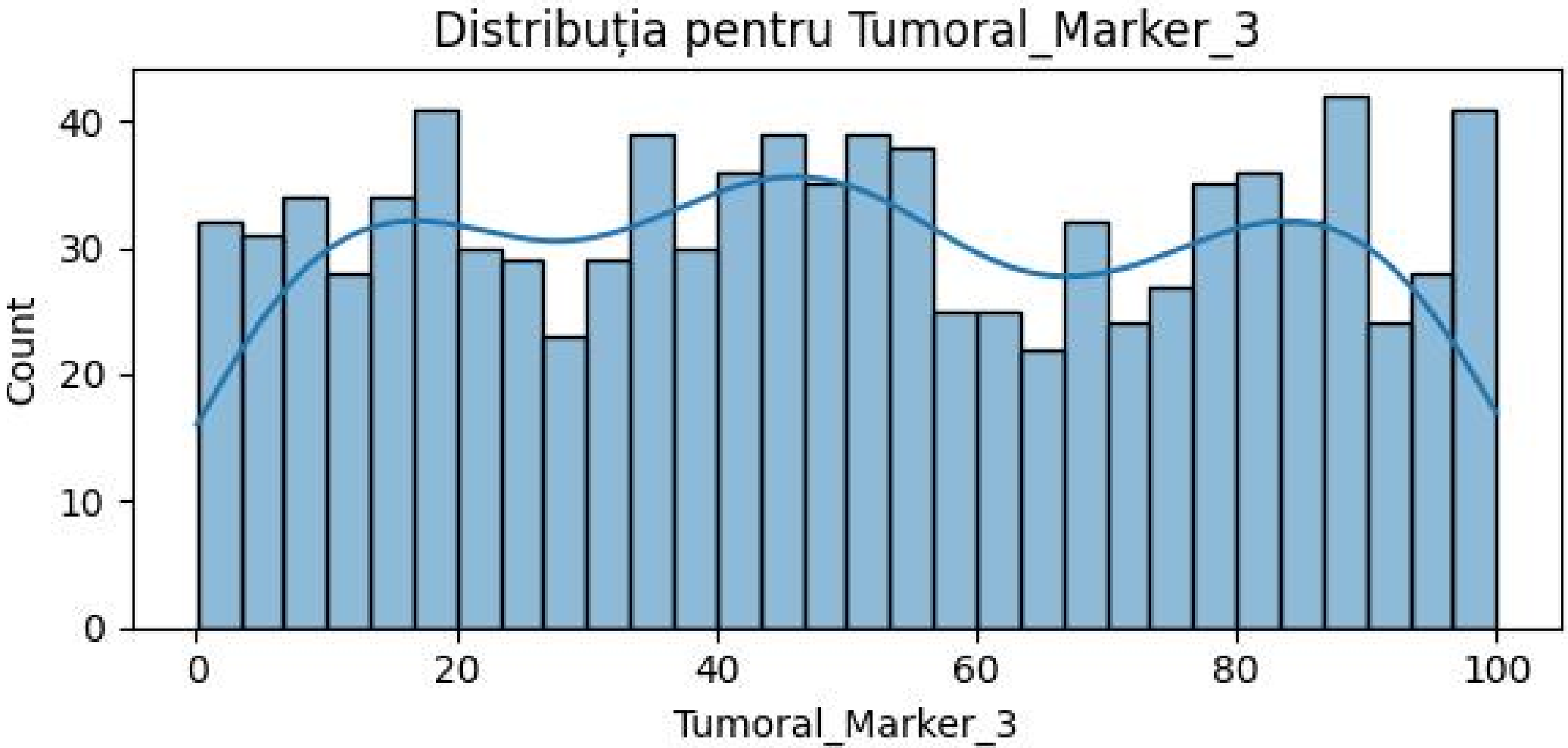

Distribuția pentru Tumoral_Marker_0
Count
0
10
20
30
40
50
0
20
40
60
80
100
Tumoral_Marker_0

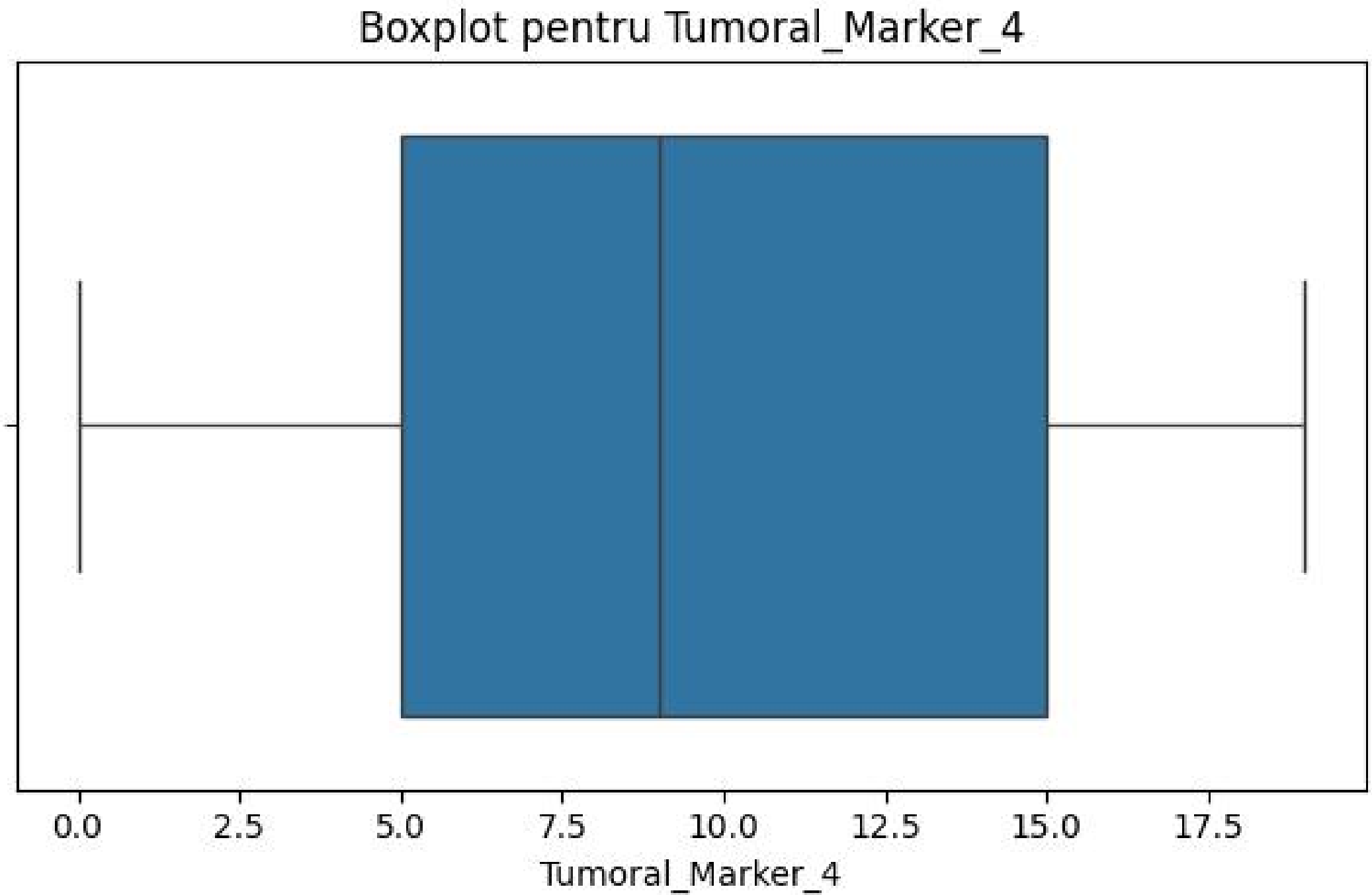

Boxplot pentru Tumoral_Marker_1
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
Tumoral_Marker_1

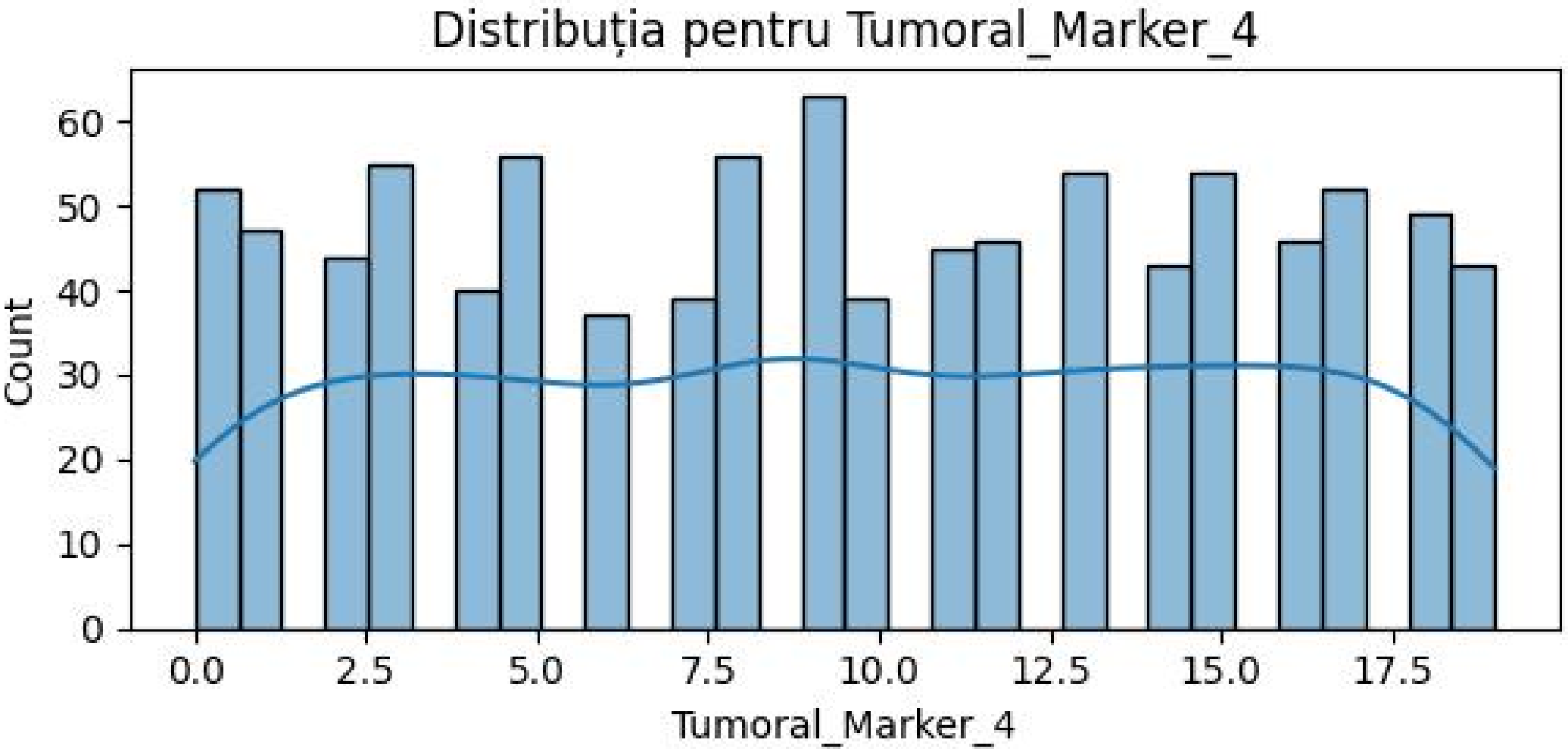

Distribuția pentru Tumoral_Marker_1
Count
0
10
20
30
40
50
60
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
Tumoral_Marker_1

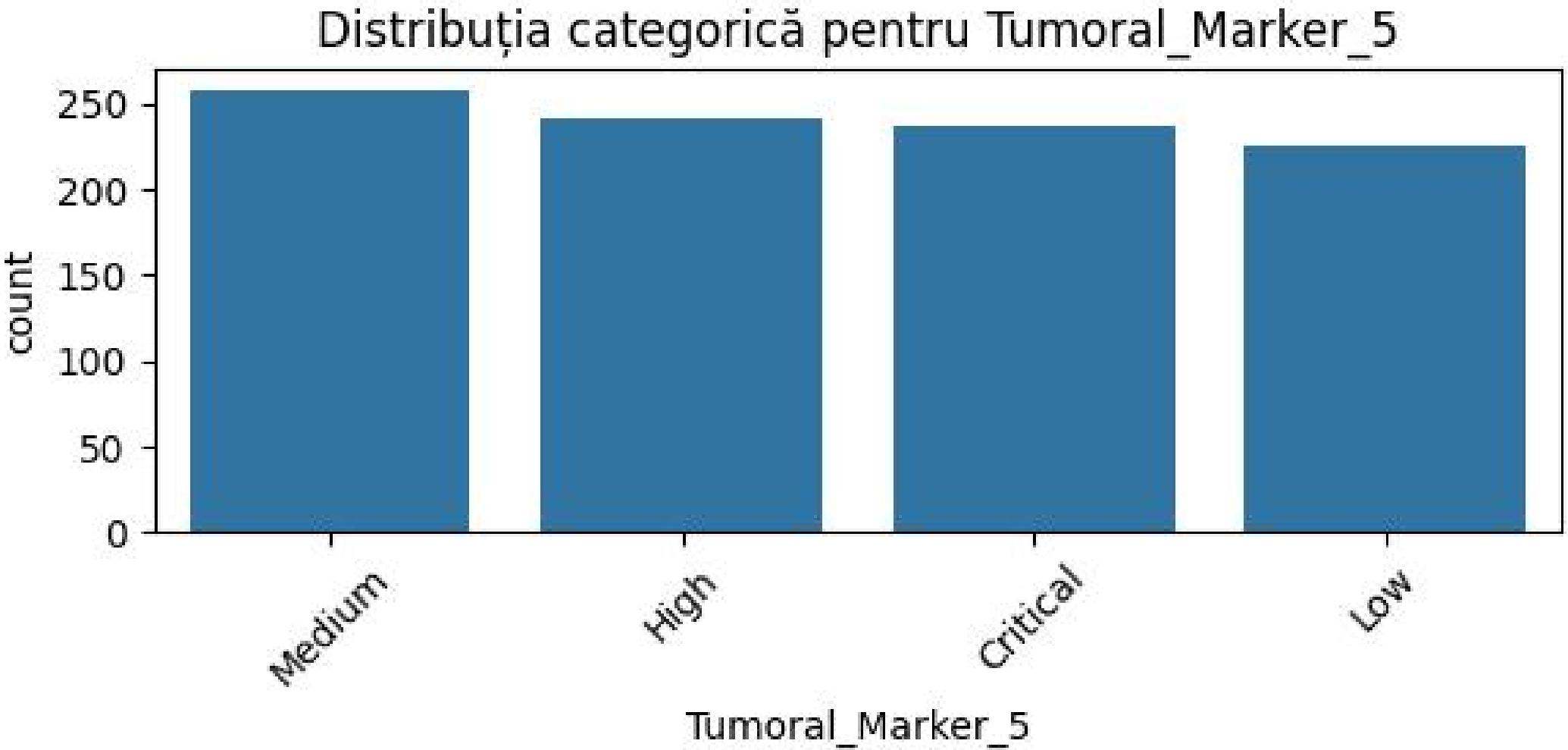

Distribuția categorică pentru Tumoral_Marker_2
count
250
200
150
100
50
0
High
Medium
Low
Critical
Tumoral_Marker_2

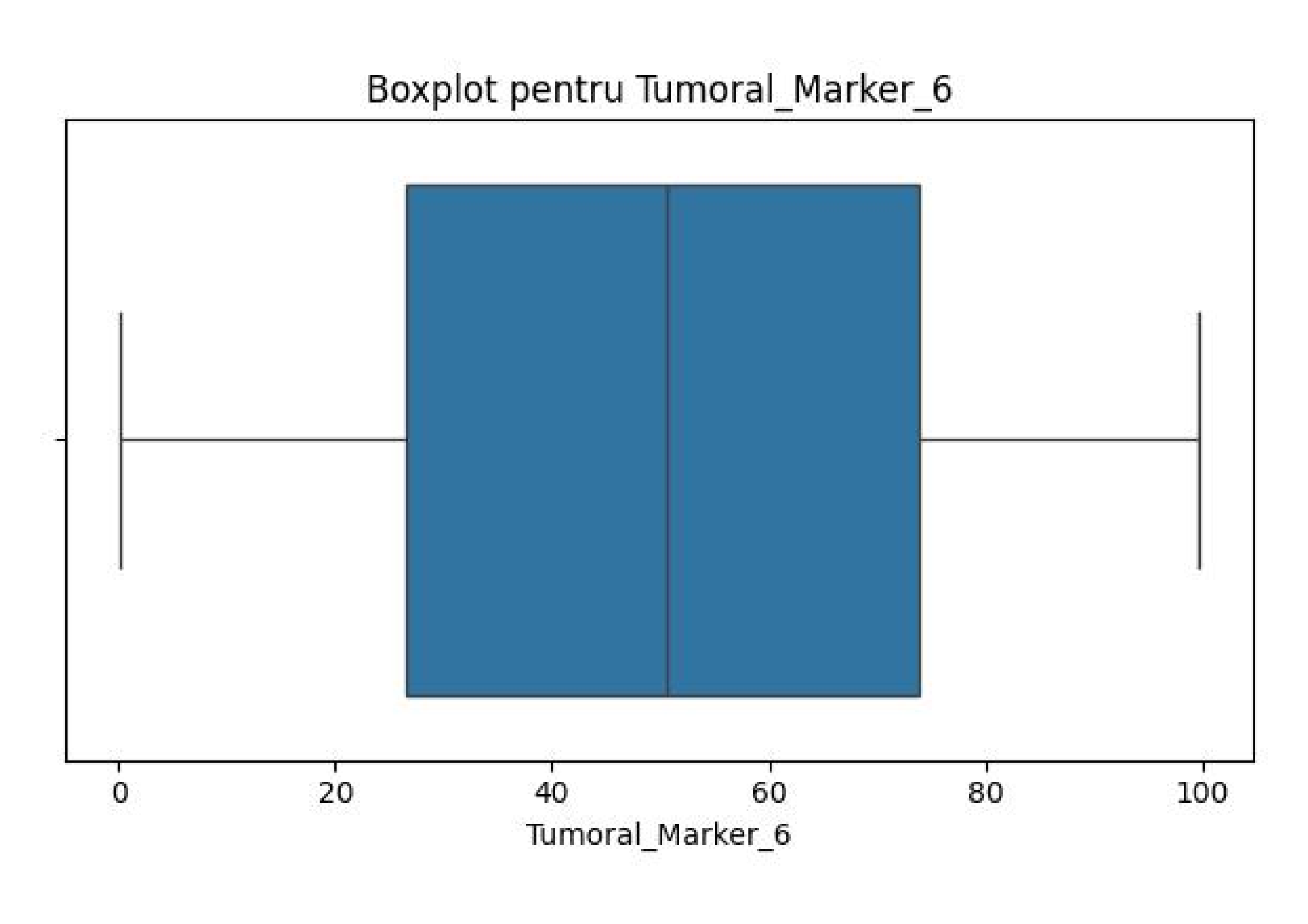

Boxplot pentru Tumoral_Marker_3
0
20
40
60
80
100
Tumoral_Marker_3

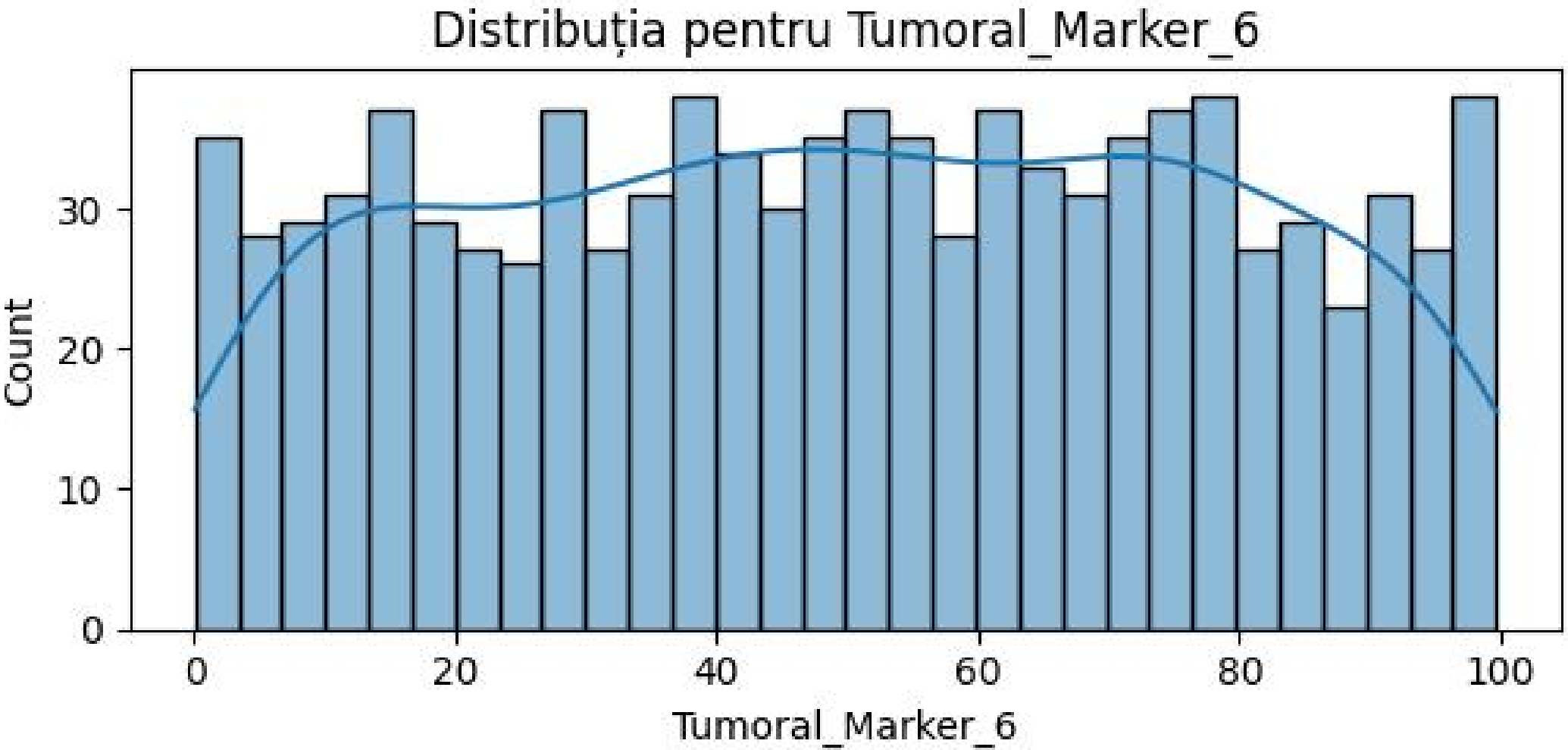

Distribuția pentru Tumoral_Marker_3
Count
0
10
20
30
40
0
20
40
60
80
100
Tumoral_Marker_3

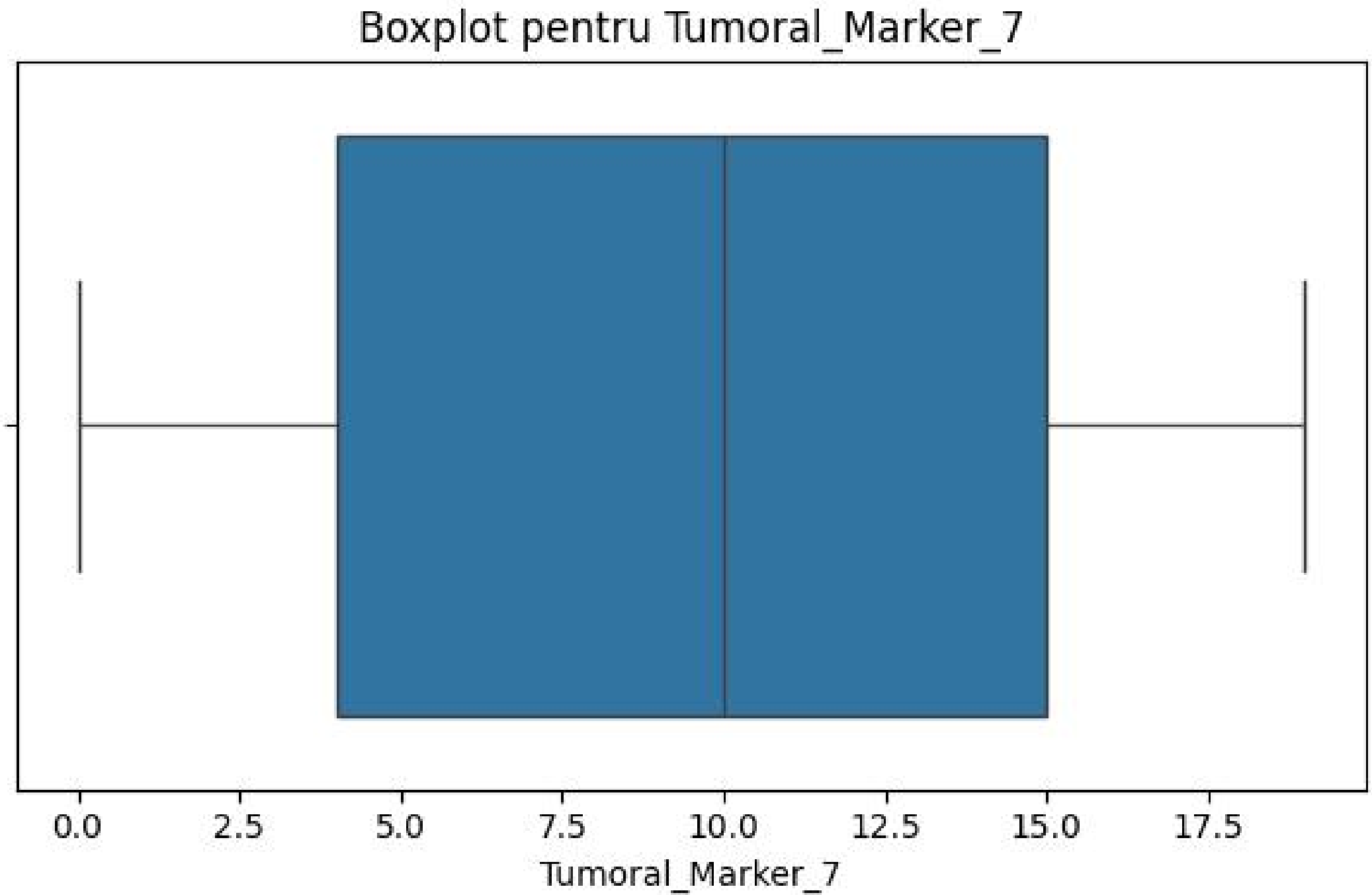

Boxplot pentru Tumoral_Marker_4
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
Tumoral_Marker_4

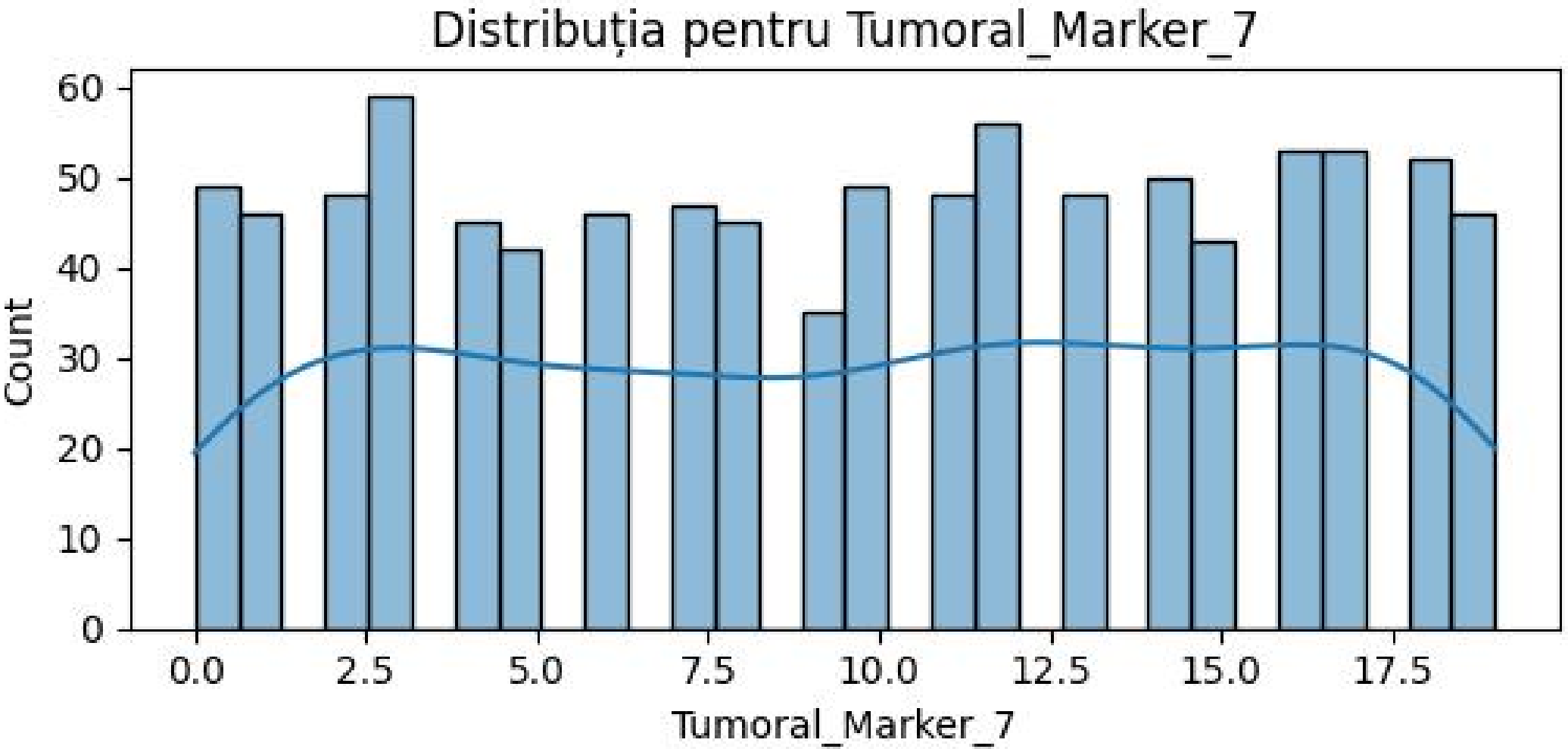

Distribuția pentru Tumoral_Marker_4
Count
0
10
20
30
40
50
60
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
Tumoral_Marker_4

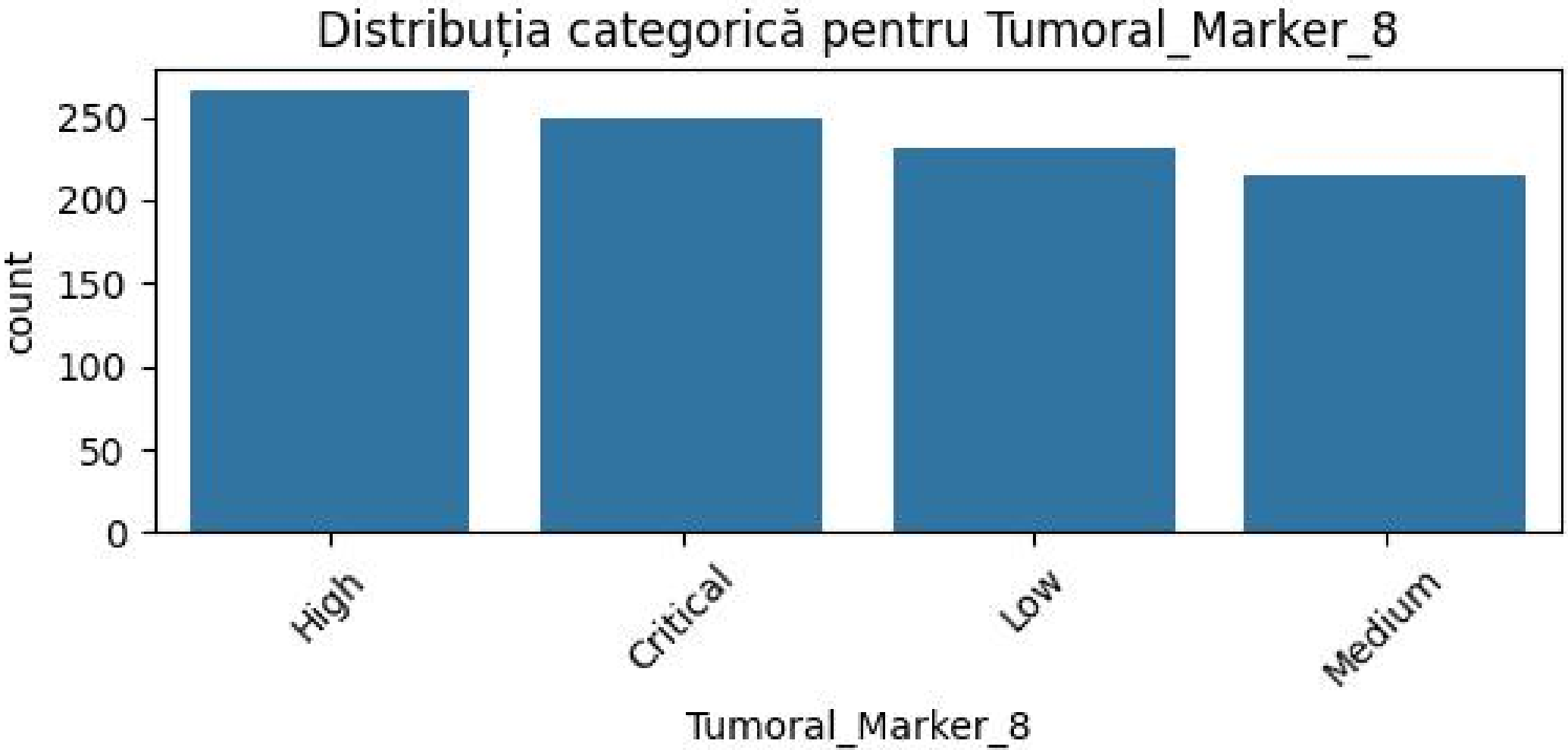

Distribuția categorică pentru Tumoral_Marker_5
count
250
200
150
100
50
0
Medium
High
Critical
Low
Tumoral_Marker_5

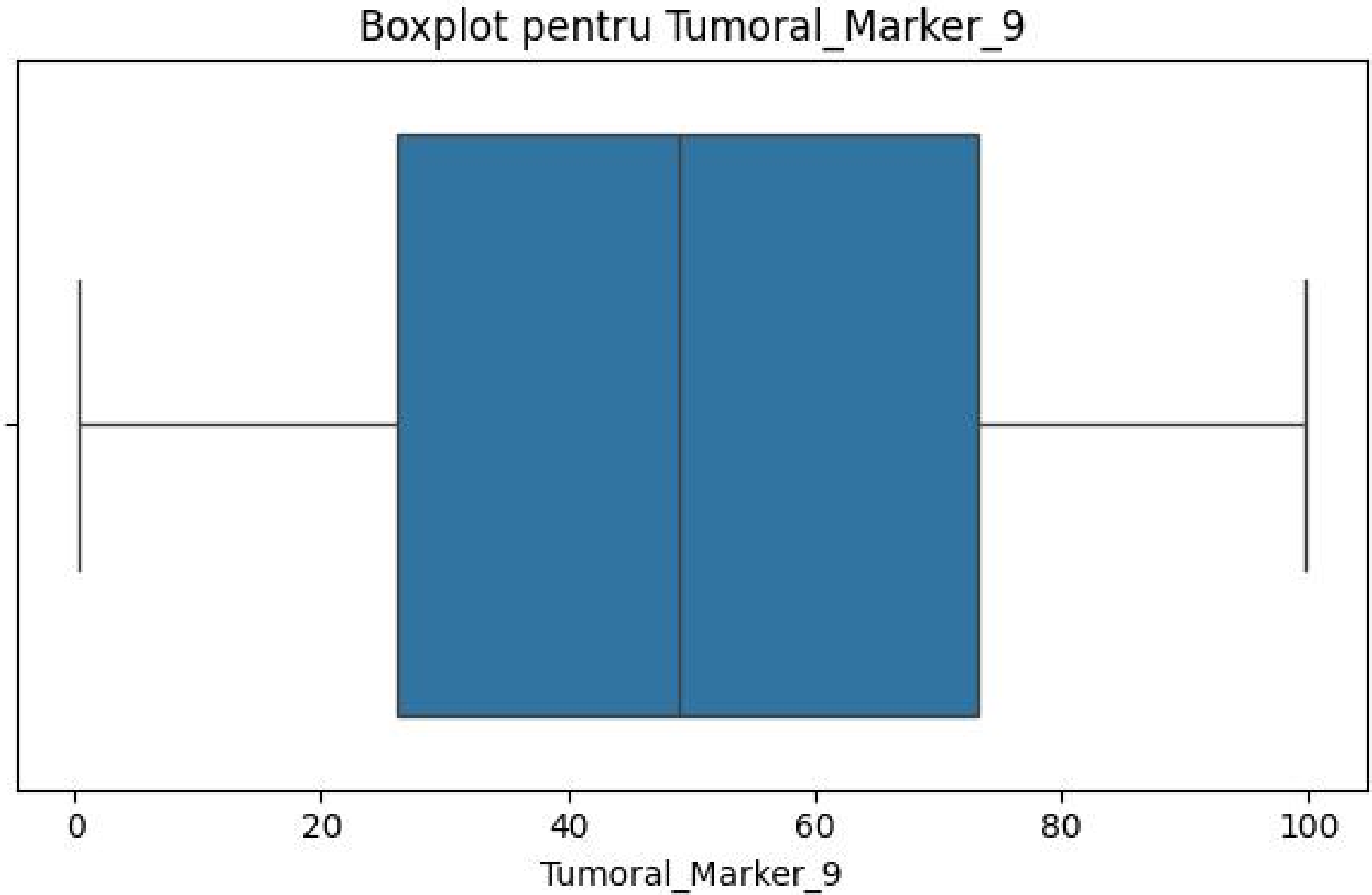

Boxplot pentru Tumoral_Marker_6
0
20
40
60
80
100
Tumoral_Marker_6

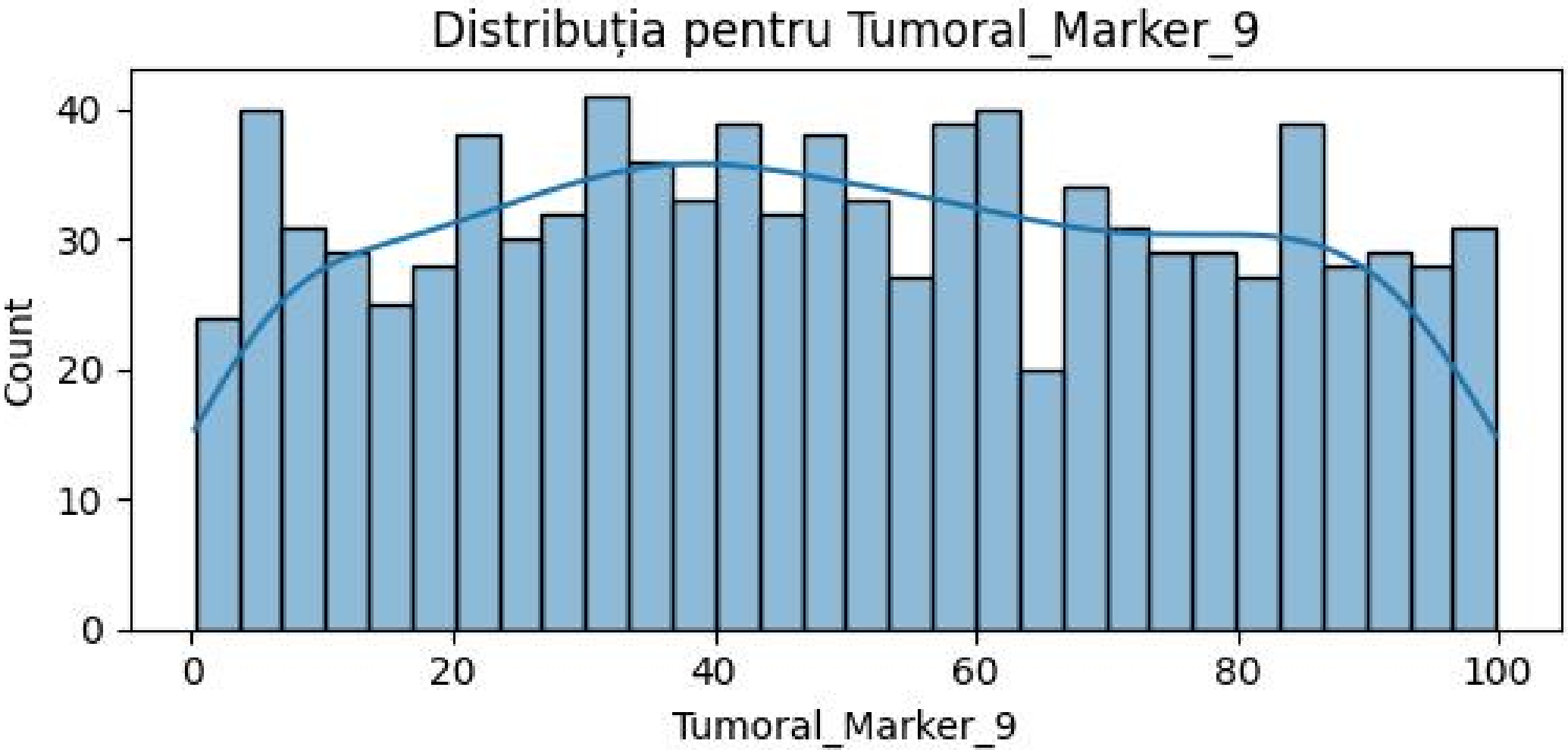

Distribuția pentru Tumoral_Marker_6
Count
0
10
20
30
0
20
40
60
80
100
Tumoral_Marker_6

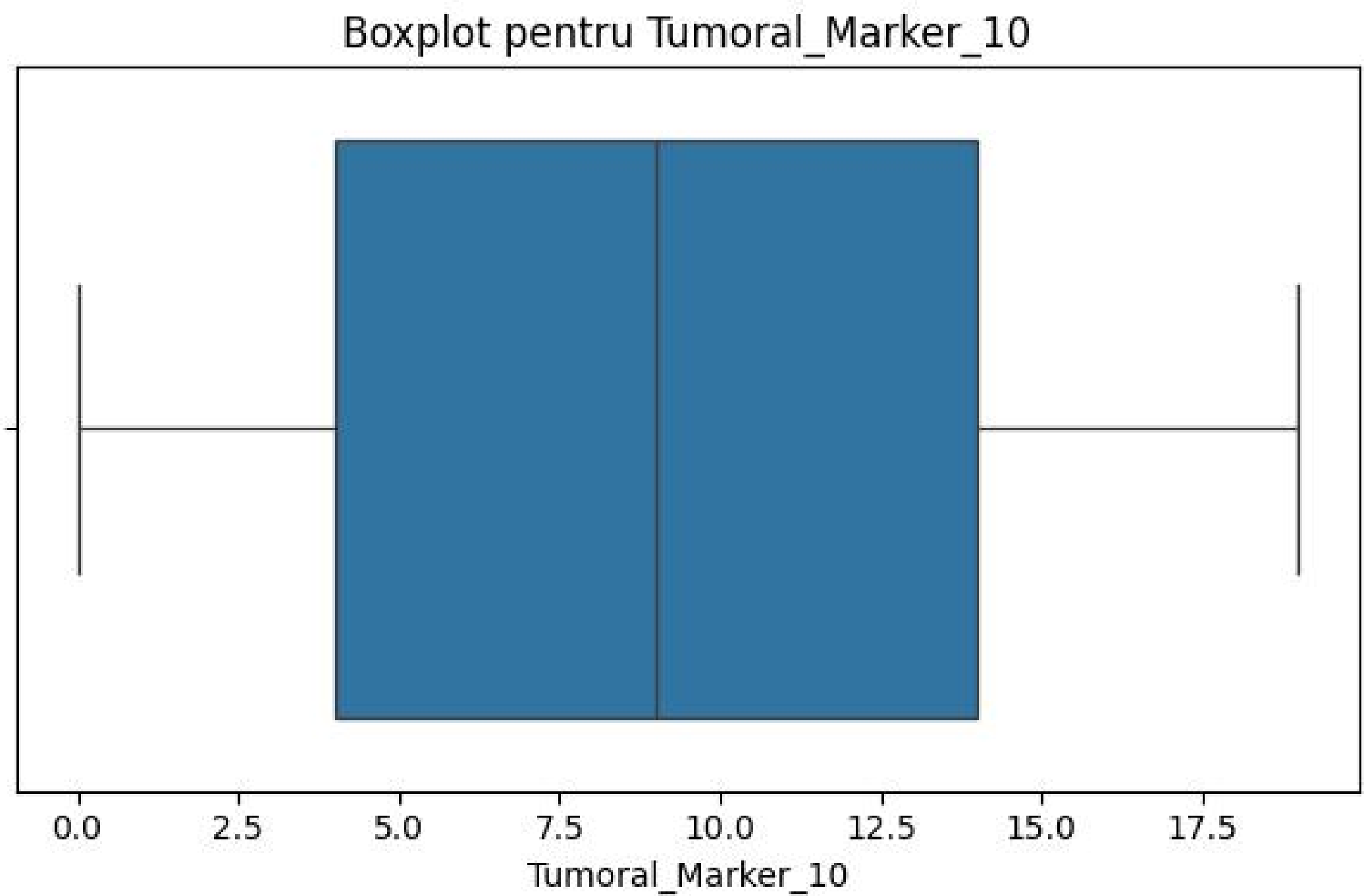

Boxplot pentru Tumoral_Marker_7
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
Tumoral_Marker_7

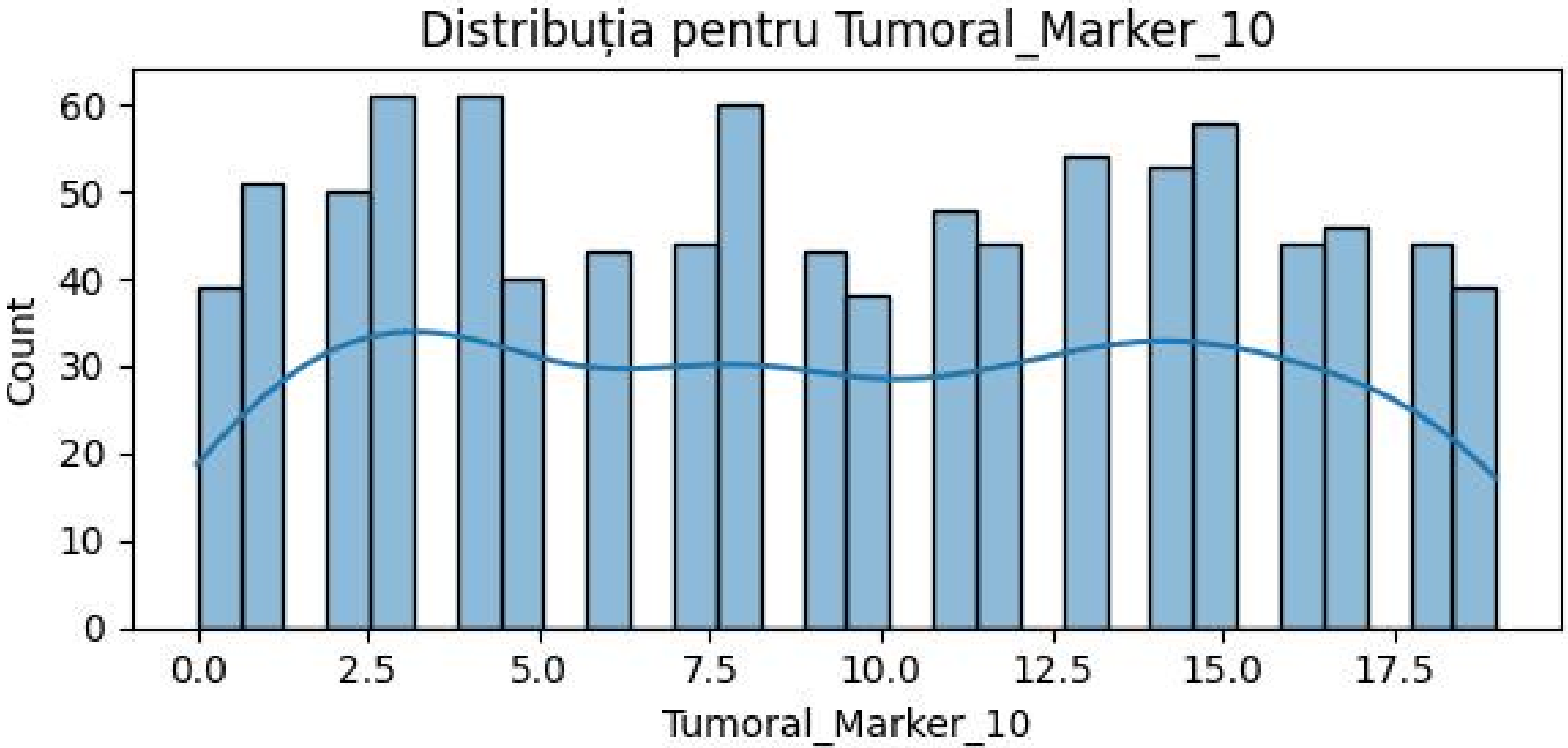

Distribuția pentru Tumoral_Marker_7
Count
60
50
40
30
20
10
0
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
Tumoral_Marker_7

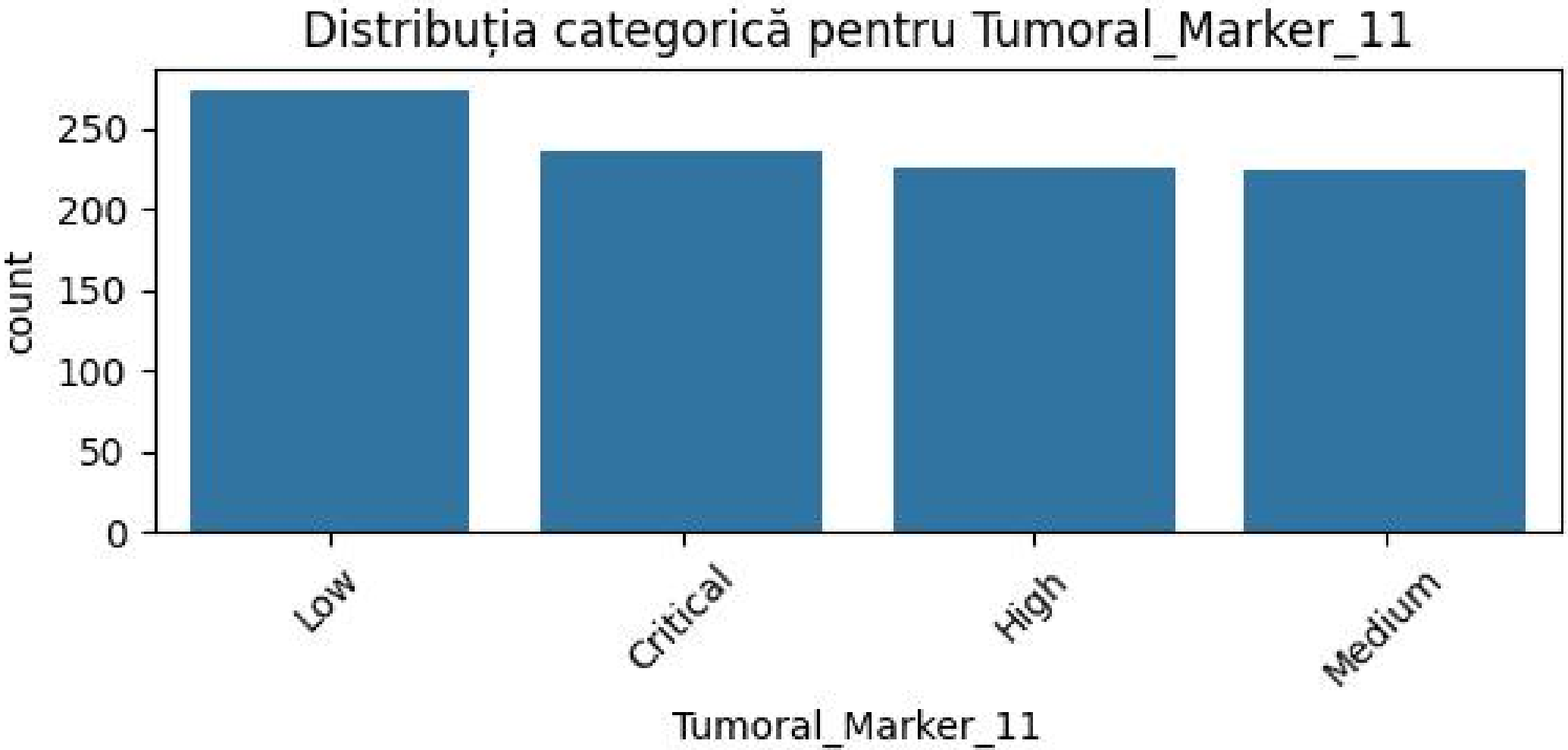

Distribuția categorică pentru Tumoral_Marker_8
count
250
200
150
100
50
0
High
Critical
Low
Medium
Tumoral_Marker_8

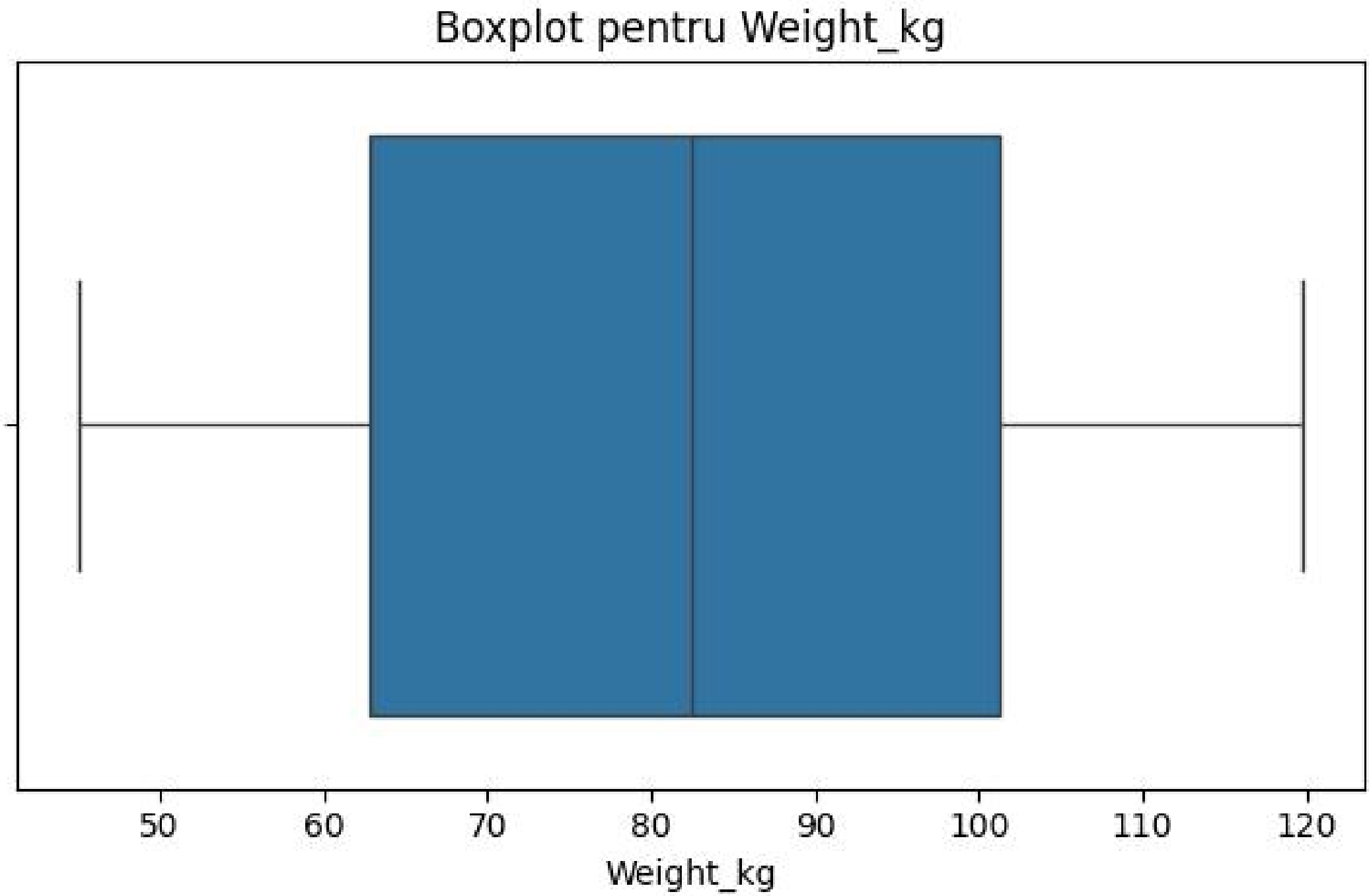

Boxplot pentru Tumoral_Marker_9
0
20
40
60
80
100
Tumoral_Marker_9

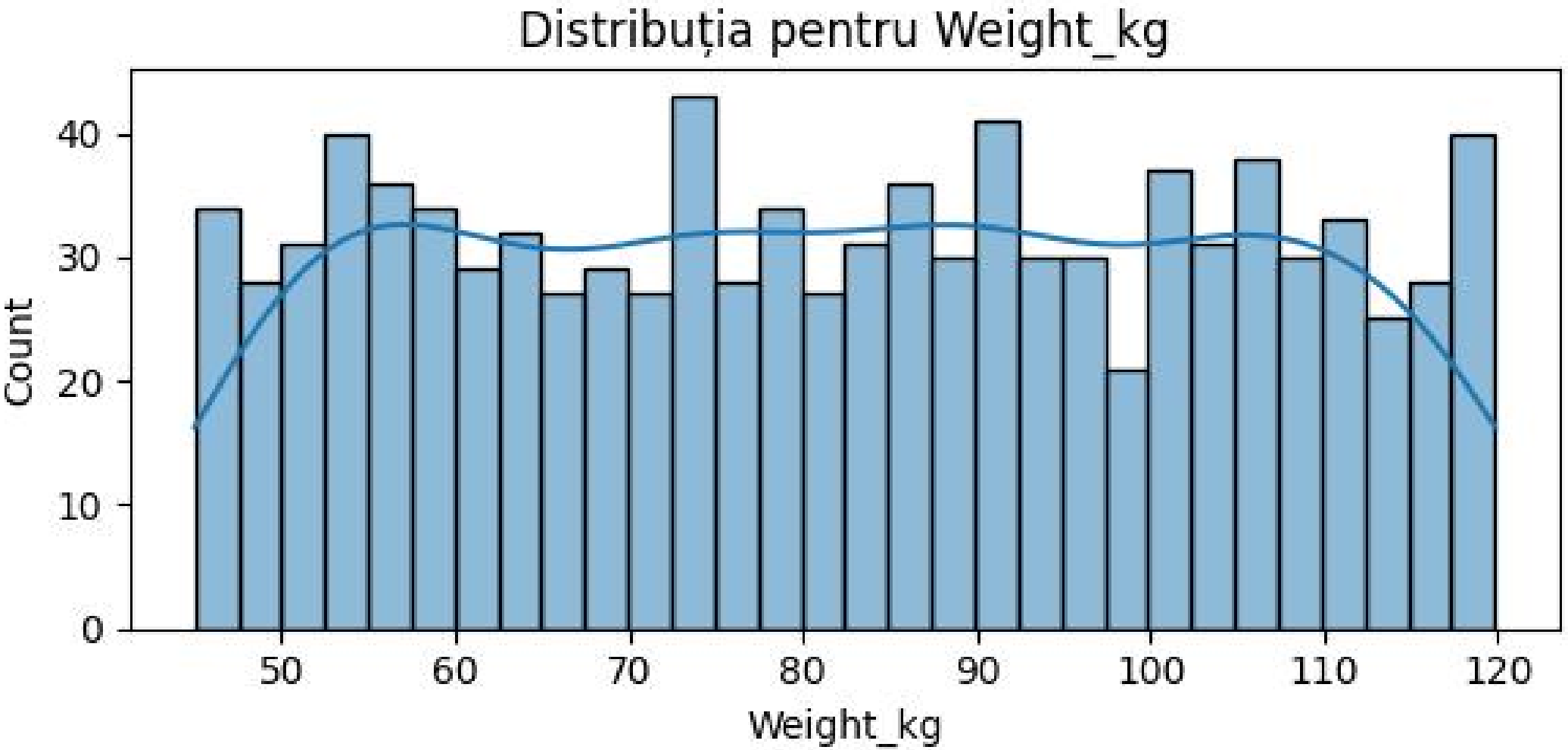

Distribuția pentru Tumoral_Marker_9
Count
0
10
20
30
40
0
20
40
60
80
100
Tumoral_Marker_9


Boxplot pentru Tumoral_Marker_10
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
Tumoral_Marker_10


Distribuția pentru Tumoral_Marker_10
Count
0
10
20
30
40
50
60
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
Tumoral_Marker_10


Distribuția categorică pentru Tumoral_Marker_11
count
250
200
150
100
50
0
Low
Critical
High
Medium
Tumoral_Marker_11


Boxplot pentru Weight_kg
50
60
70
80
90
100
110
120
Weight_kg


Distribuția pentru Weight_kg
Count
0
10
20
30
40
50
60
70
80
90
100
110
120
Weight_kg