Tema PCLP3
Partea II
Nume: Solomon Stefan
Coleg: Carp Andrei Costin
Grupa: 312 CA
Github: https://github.com/krpandrei05/Tema1_PCLP3

1. Prelucrarea datelor

1. a) Encodarea valorilor categorice: Datele contin coloane cu valori de tip text (aplicatie, daca clientul utilizeaza cupoane, ce dispozitiv foloseste, metoda de plata). Am folosit LabelEncoder pentru coloane binare (aplicatia, cupoane_de_reducere) si OneHotEncoding pentru coloane cu mai multe valori posibile (dispozitiv, metoda_de_plata).

1. b) Normalizarea si standardizarea valorilor numerice: Normalizare cu MinMaxScaler (aduce valorile intre 0 si 1) si Standardizare cu StandardScaler (transforma datele intr-o distributie cu media 0 si deviatie standard 1).
Coloane prelucrate: valoare_comanda_medie, timp, frecventa_lunara, varsta, recenzii.
Datele prelucrate au fost salvate in doua fisiere separate: train_encodat.csv (pentru antrenare) si test_encodat.csv (pentru testare).

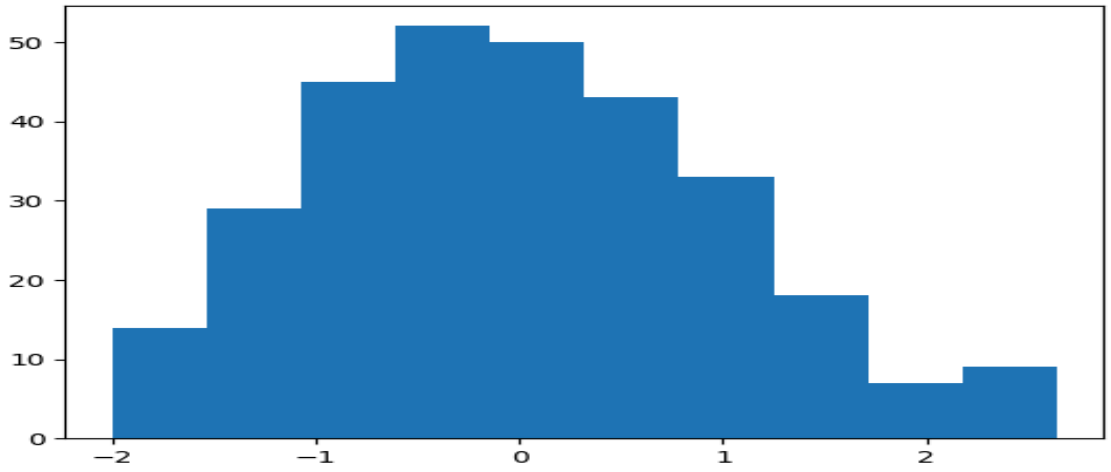
2. Analiza exploratie a datelor (EDA complex) dupa aplicarea prelucrarilor

2. a) Analiza valorilor lipsa: Am verificat fiecare coloana pentru a vedea daca exista date lipsa. Am calculat valoarea absoluta si procentul valorilor lipsa. Rezultat: Nu s-au identificat valori lipsa in seturile de date prelucrate.

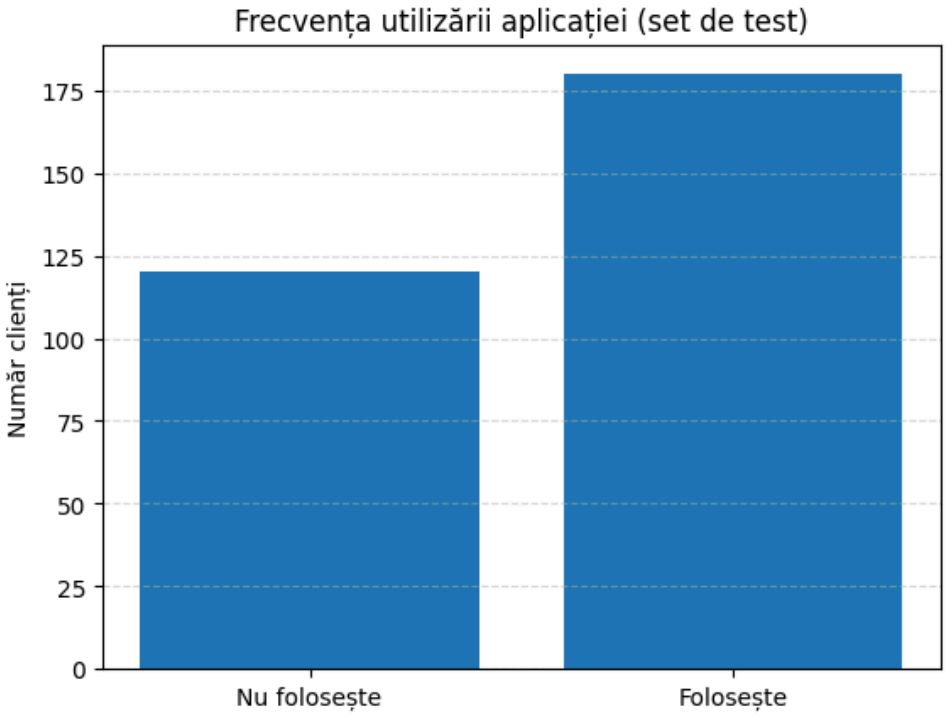
2. b) Statistici descriptive: Am generat statistici pentru fiecare coloana: media, deviatia standard, valorile minime si maxime etc.

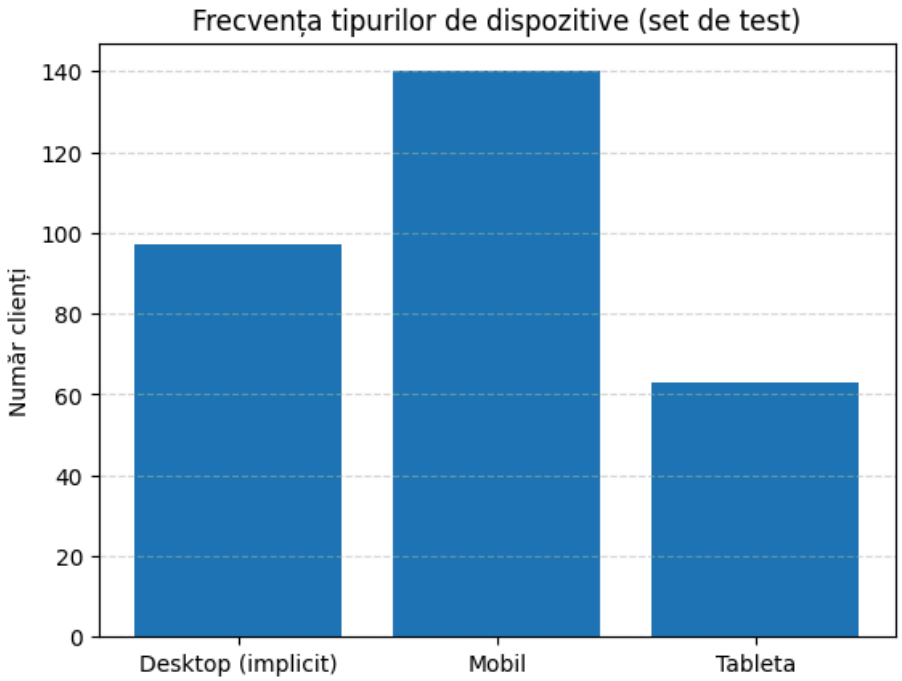
2. c) Analiza distributiei variabilelor: Pentru fiecare coloana am folosit histograme pentru valoare_comanda_medie, timp, frecventa_lunara, etc. si am folosit grafice bara pentru variabile care sunt categorice la baza (aplicatia, dispozitiv, cupoane_de_reducere). Am reconstruit si valorile eliminate prin

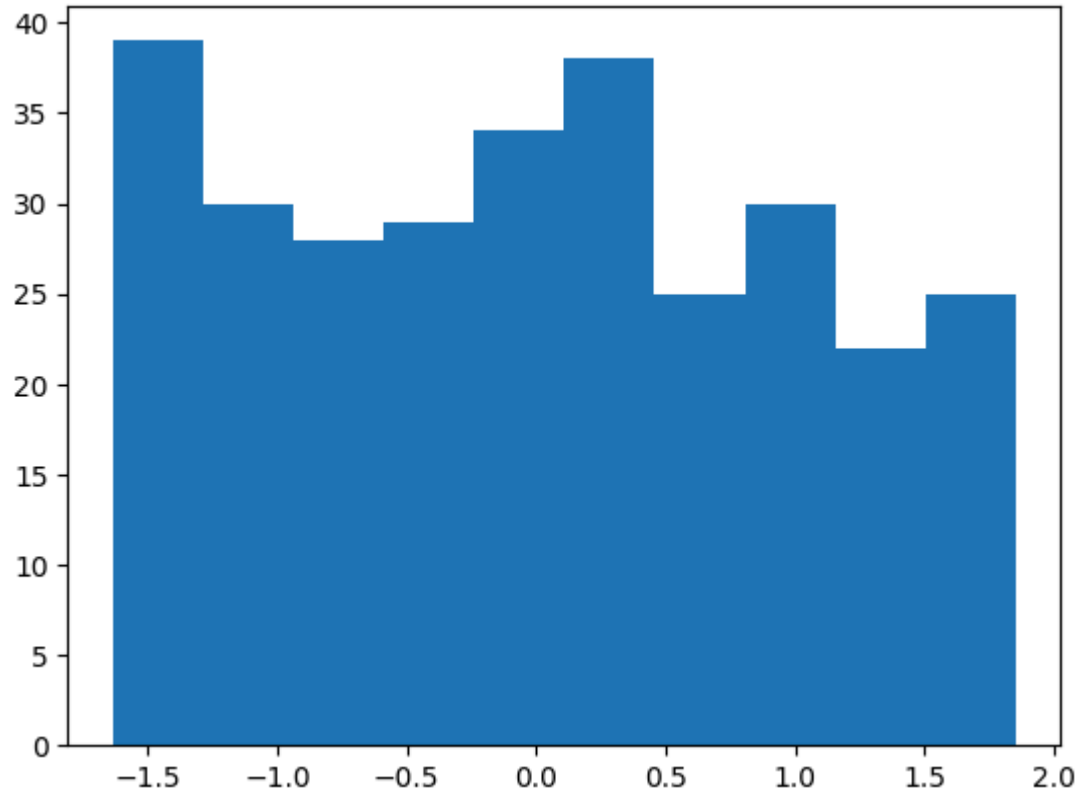
One-Hot Encoding (ex: dispozitiv_desktop).



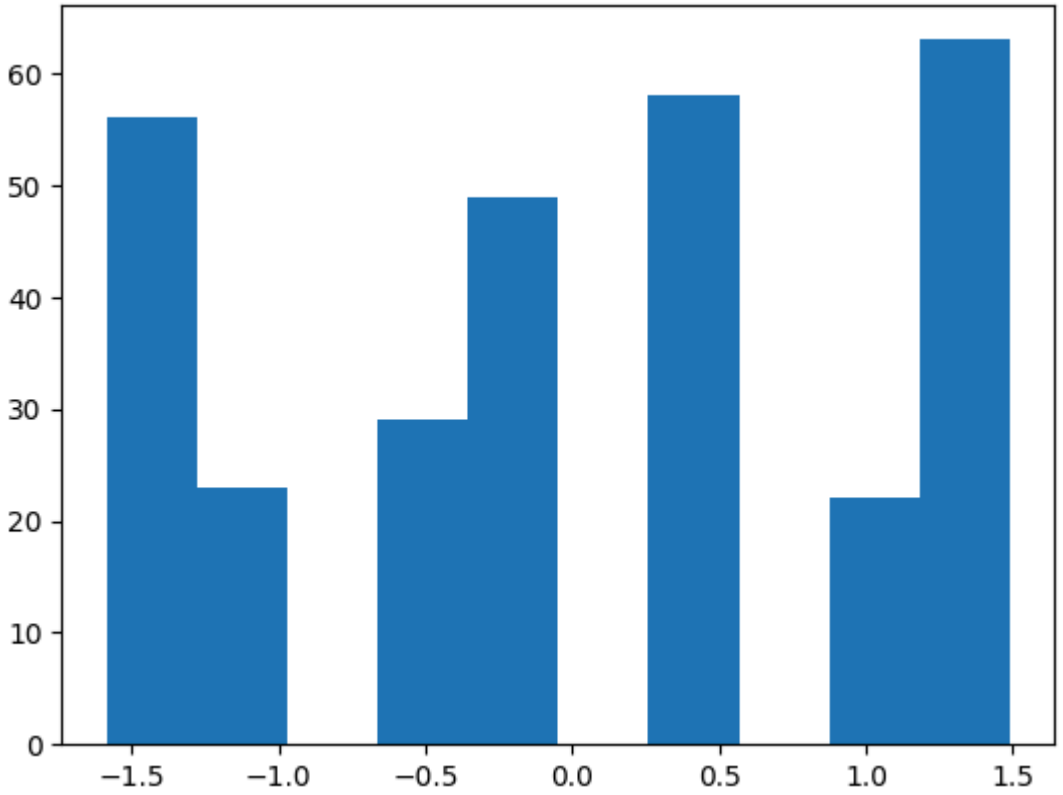
Valoarea medie a unei comenzi



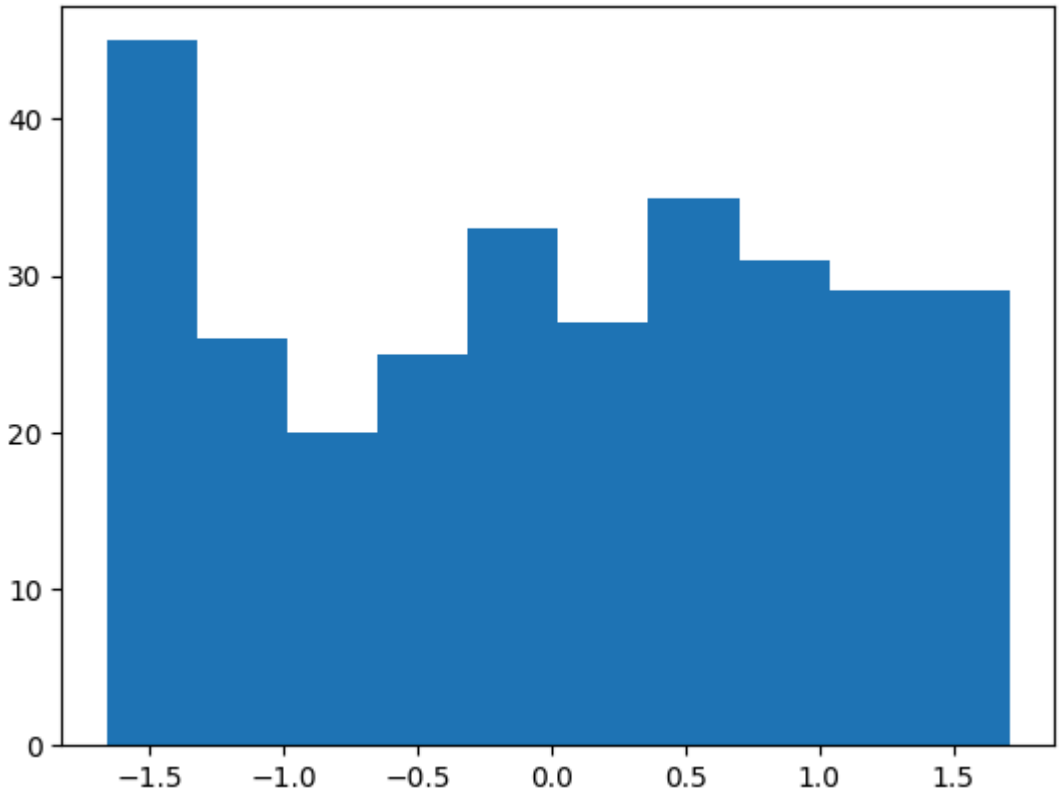




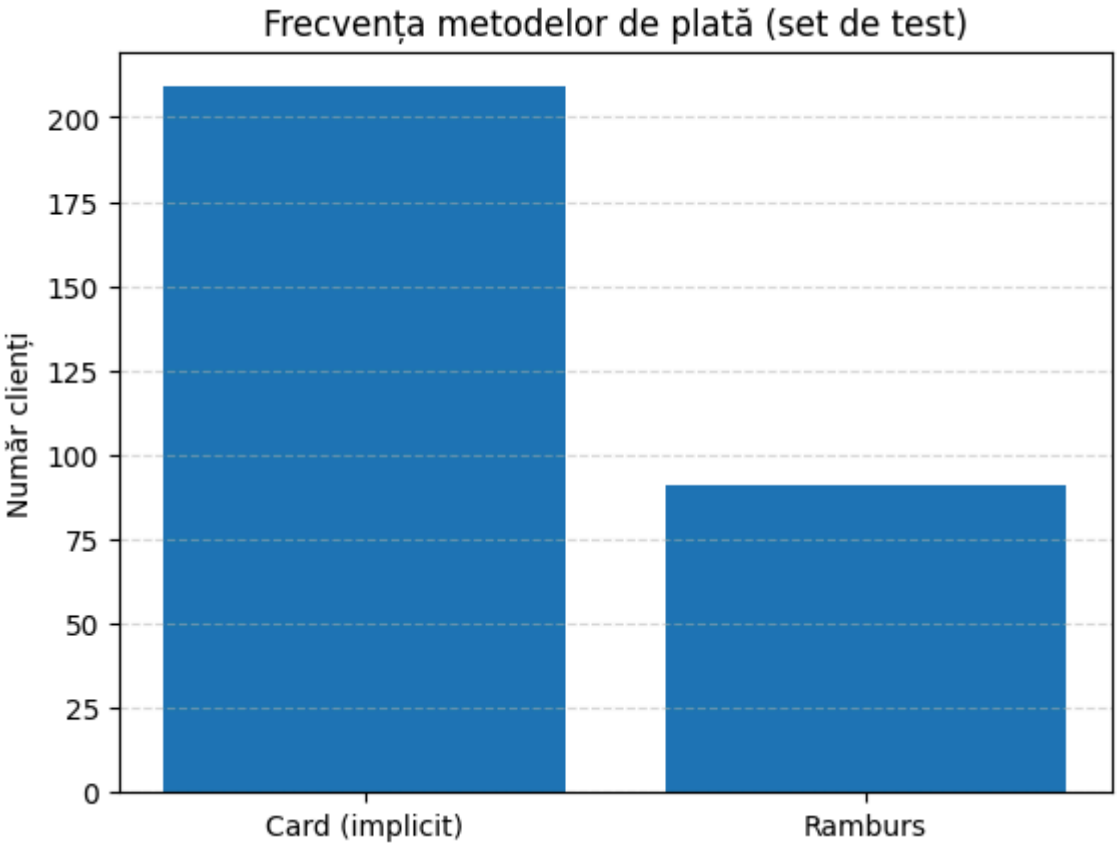
Cat timp petrece pe magazin

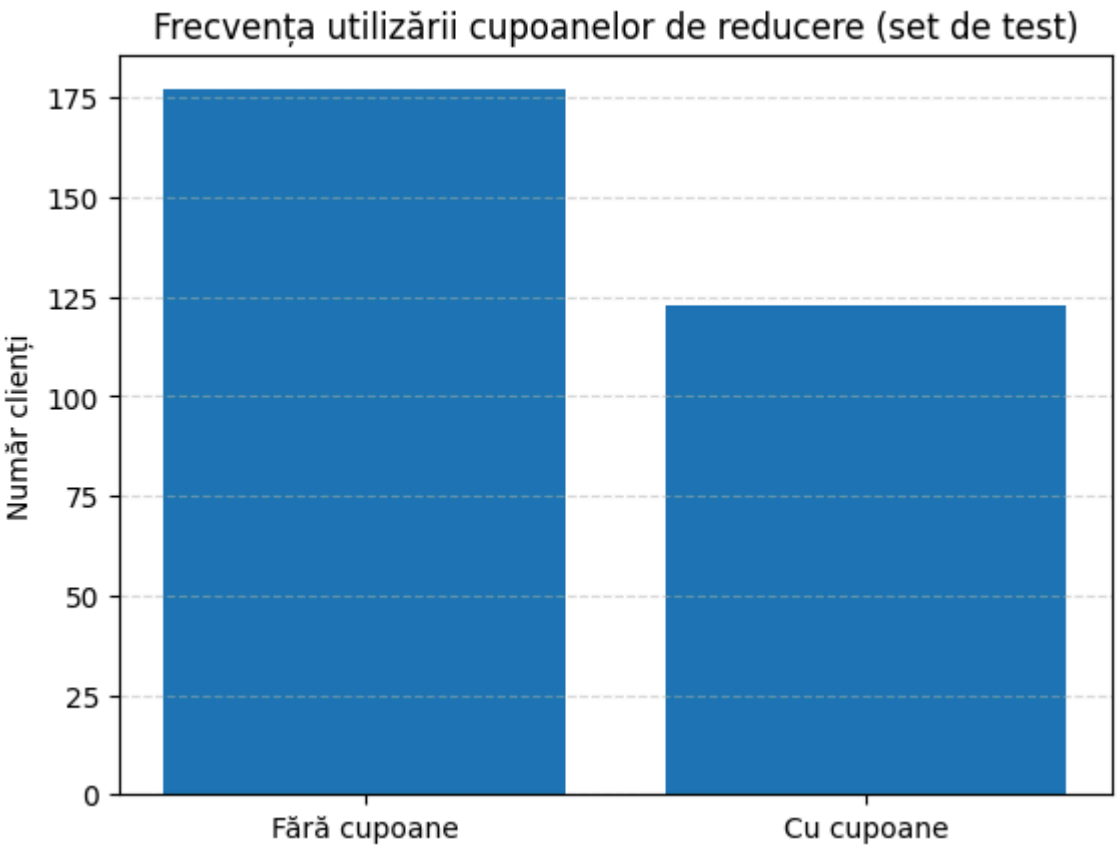


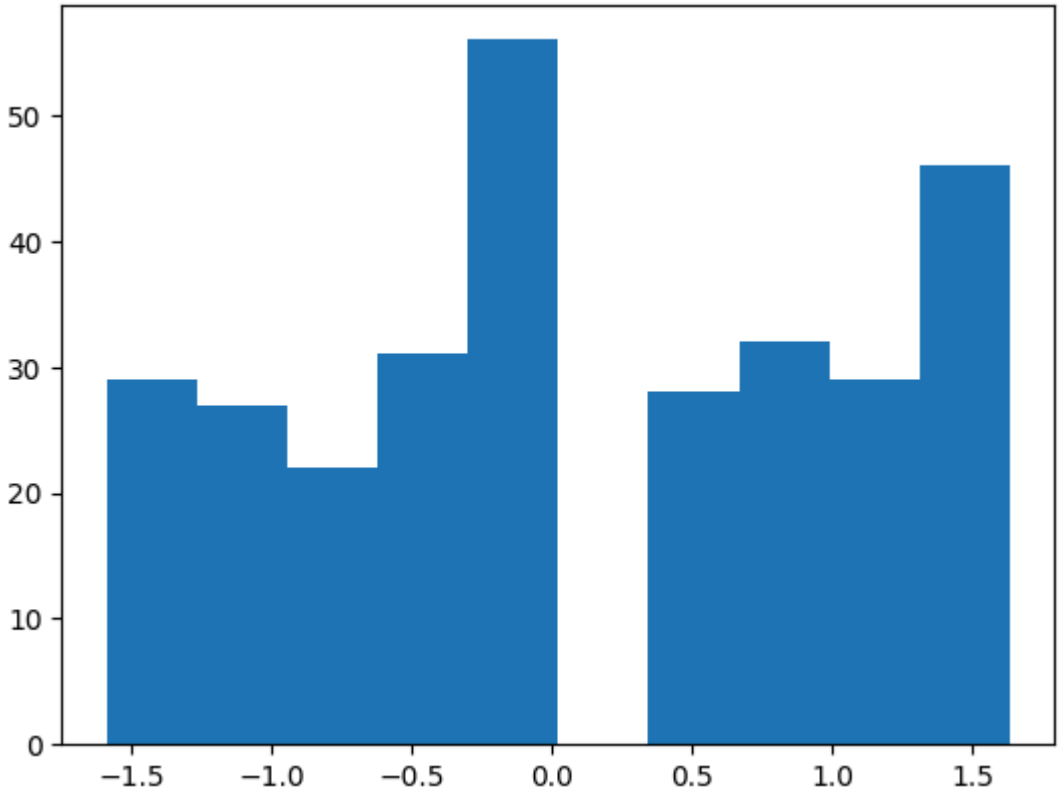
Frecventa lunara



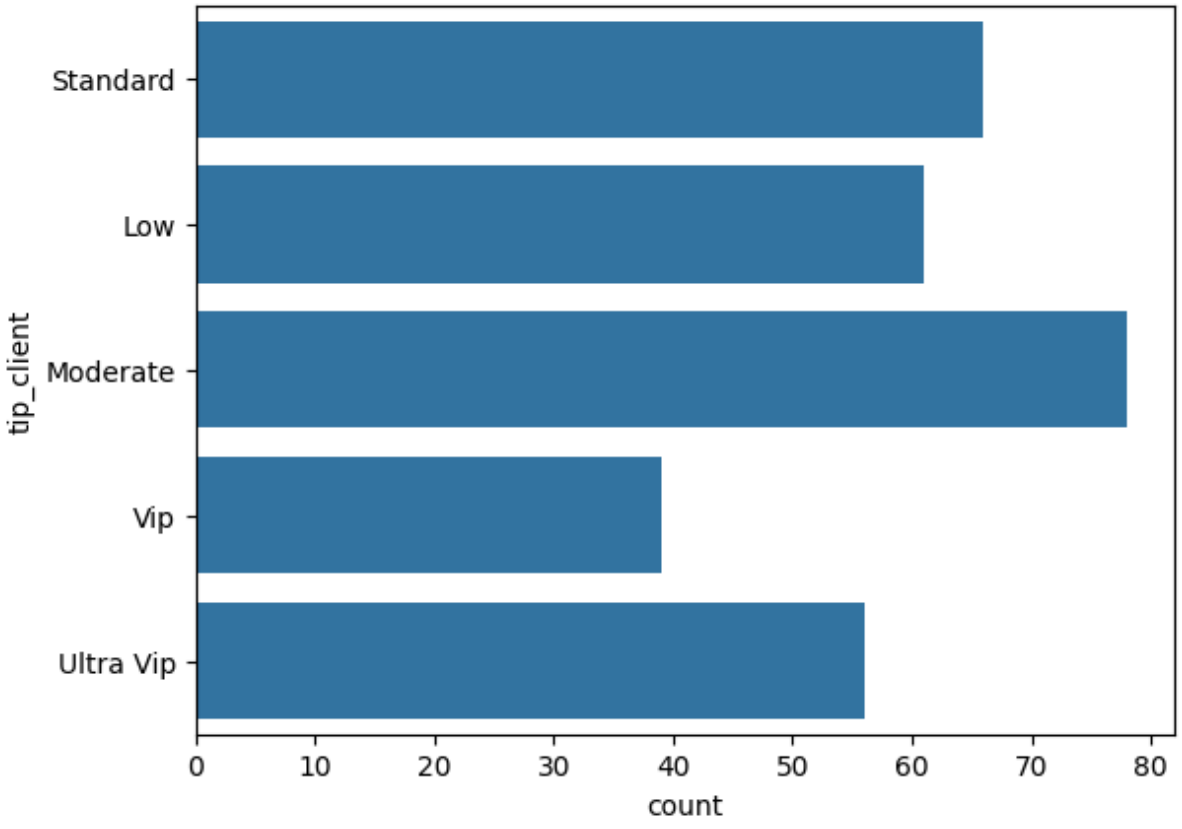
Varsta


Frecvența metodelor de plată (set de test)
Număr clienți
0
25
50
75
100
125
150
175
200
Card (implicit)
Ramburs


Frecvența utilizării cupoanelor de reducere (set de test)
Număr clienți
0
25
50
75
100
125
150
175
Fără cupoane
Cu cupoane

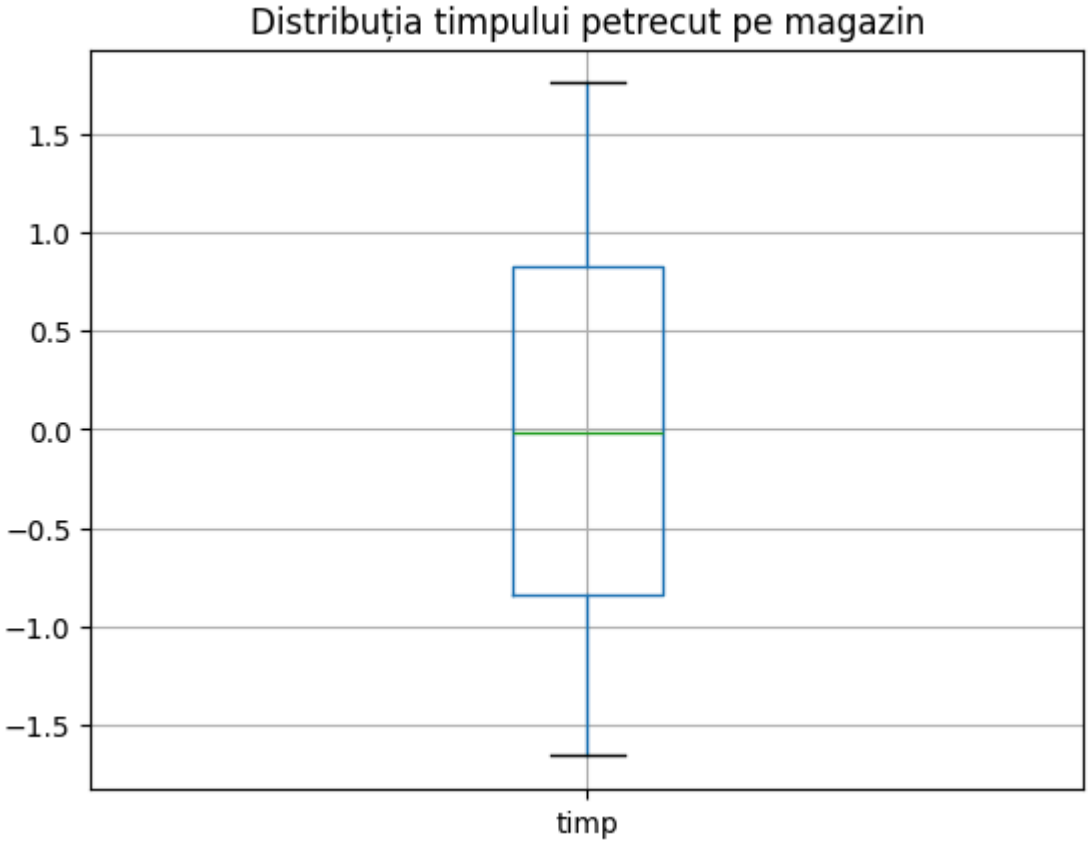


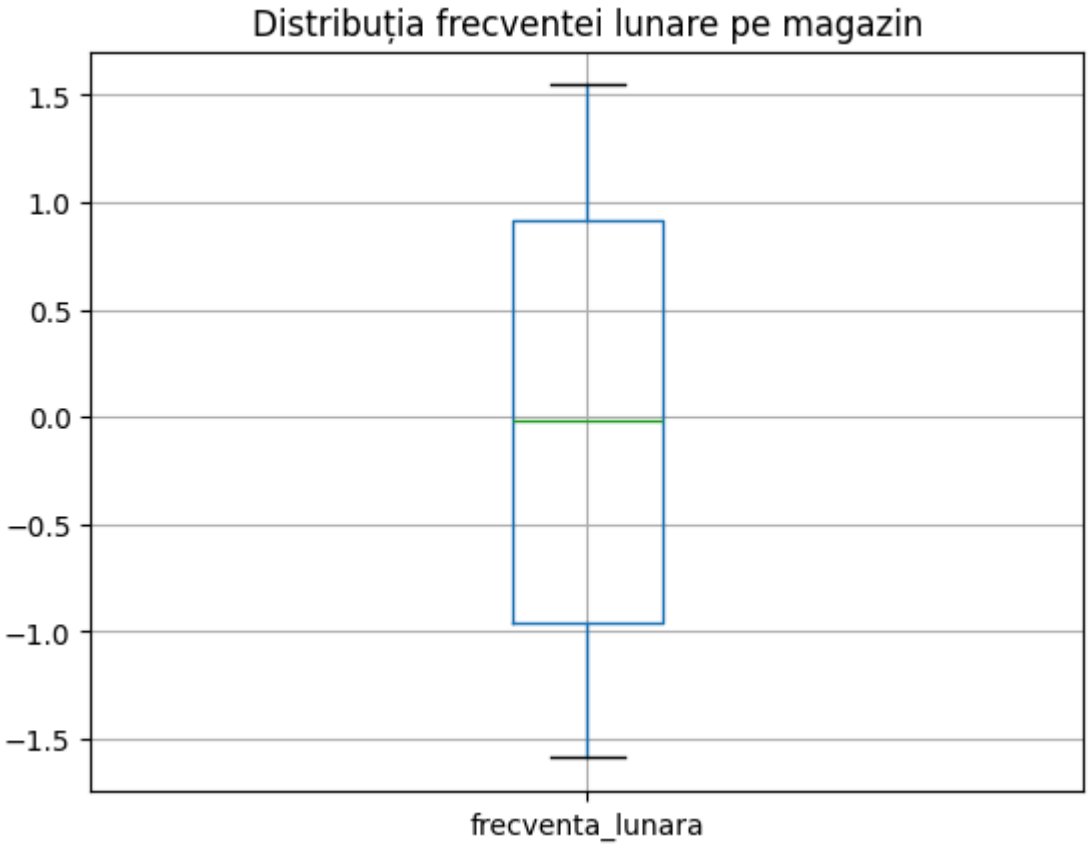
Cate recenzii acorda

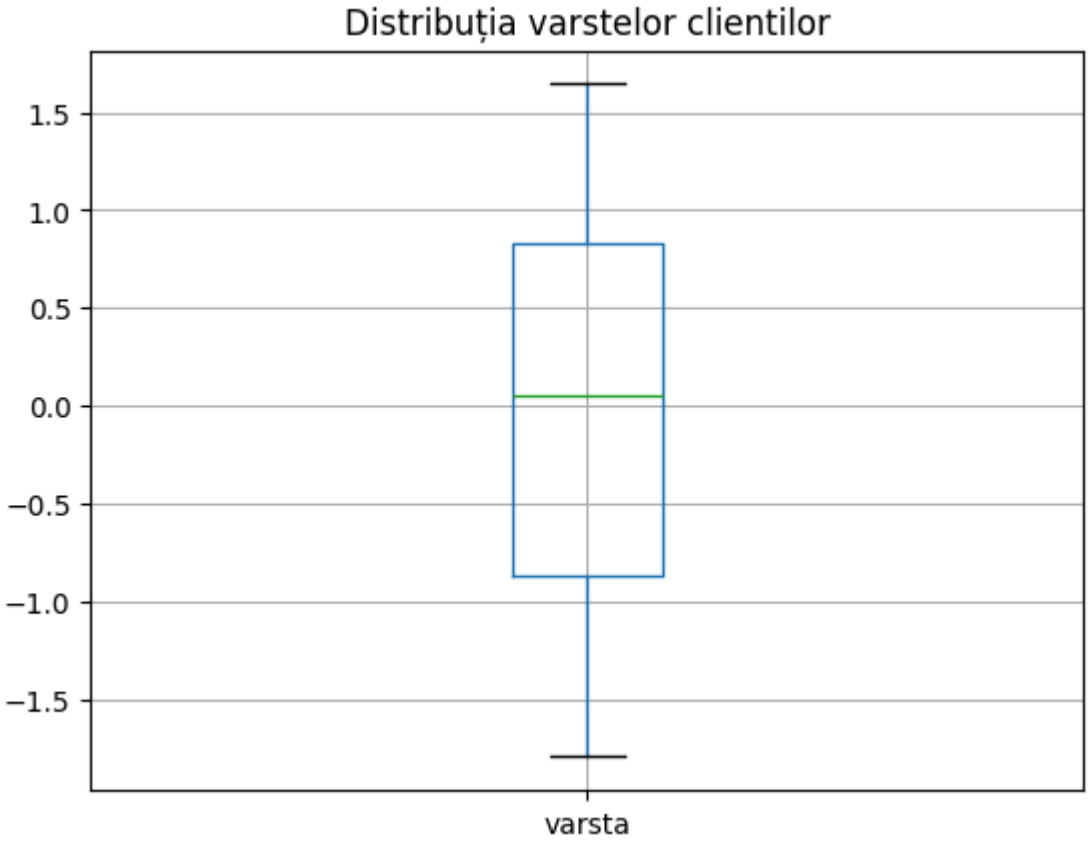


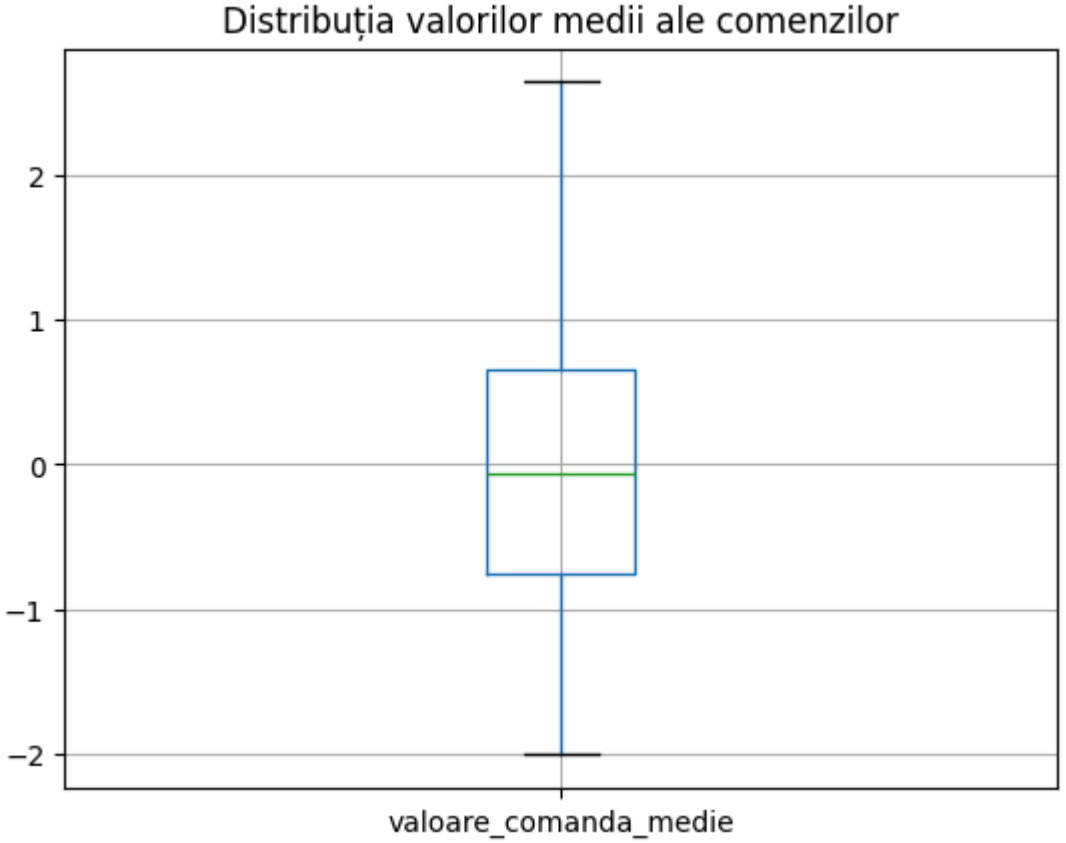
Tipuri de clienti

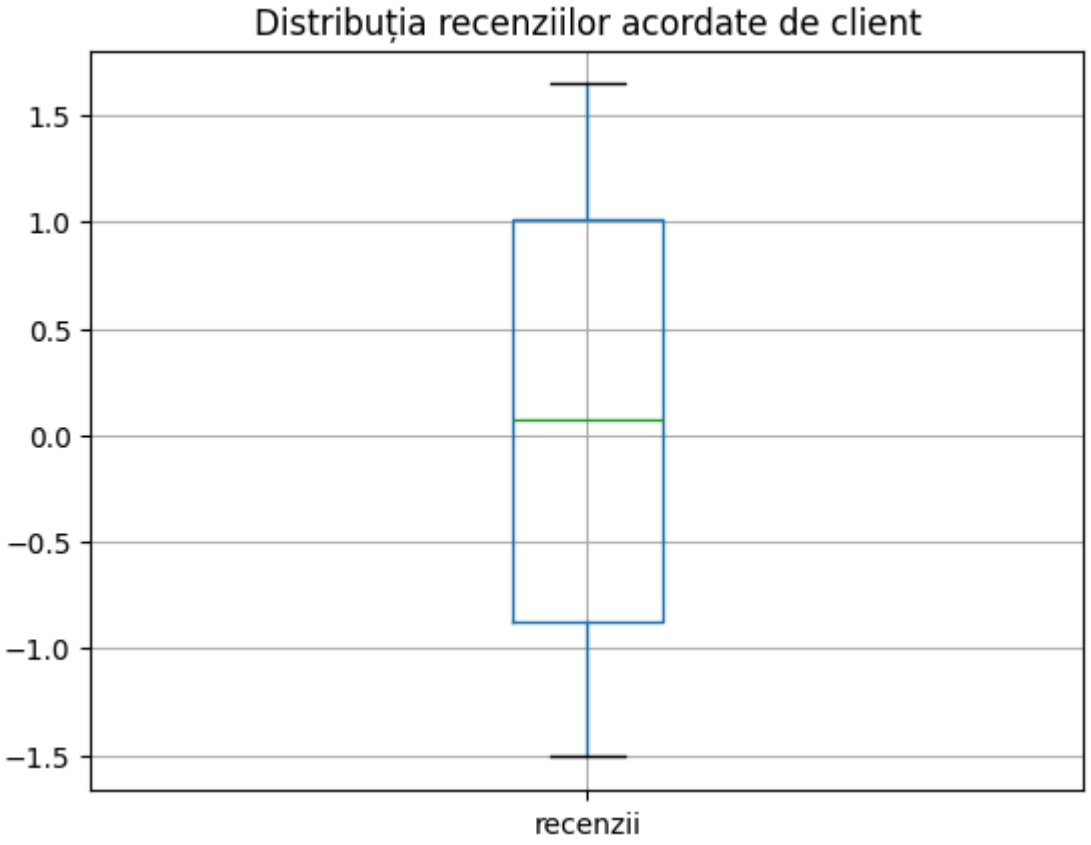
2. d) Detectarea outlierilor: Am utilizat boxplot-uri pentru a vizualiza distributia fiecarei variabile numerice. Am analizat valoare_comanda_medie, timp, frecventa_lunara, varsta, recenzii.

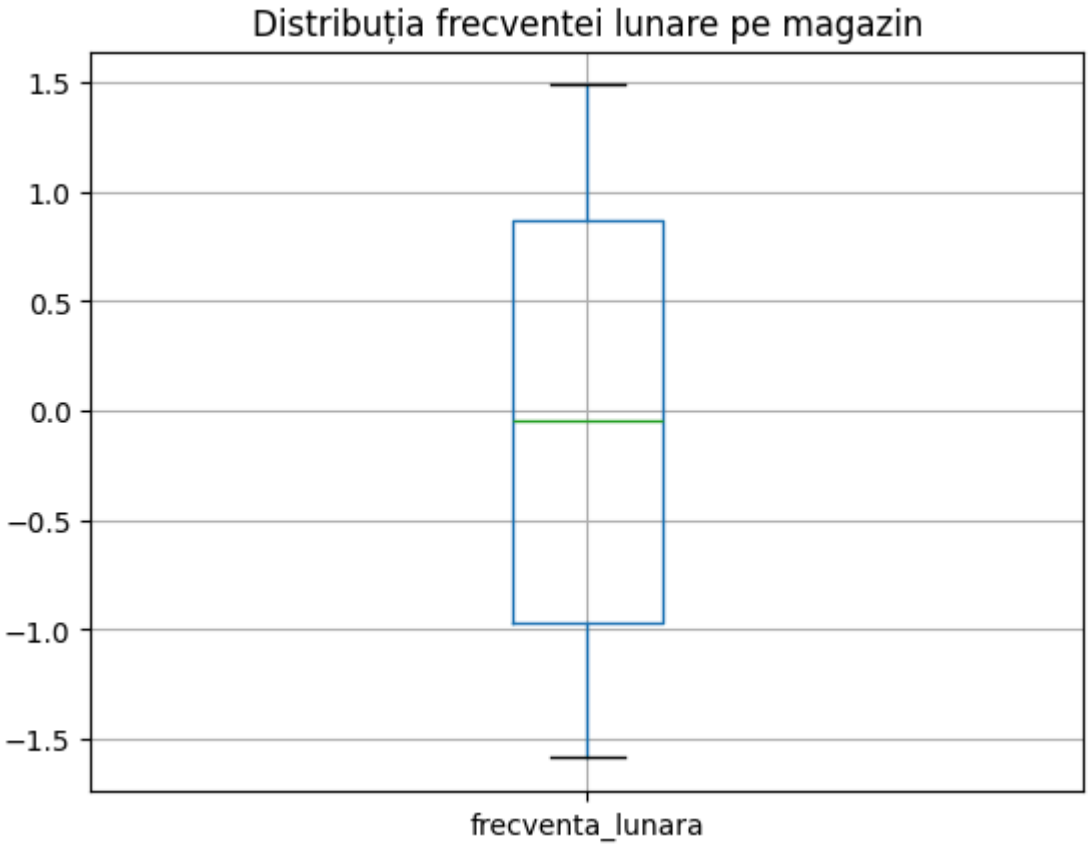


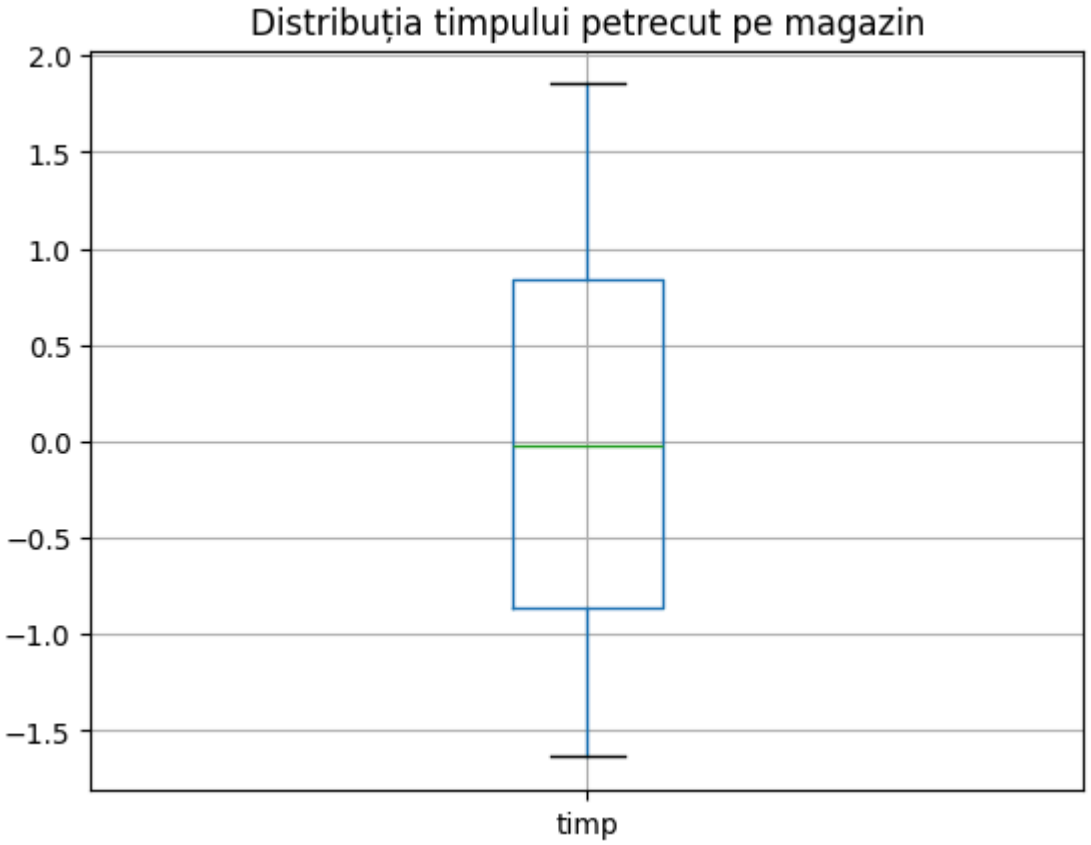

Distribuția frecventei lunare pe magazin
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
frecventa_lunara

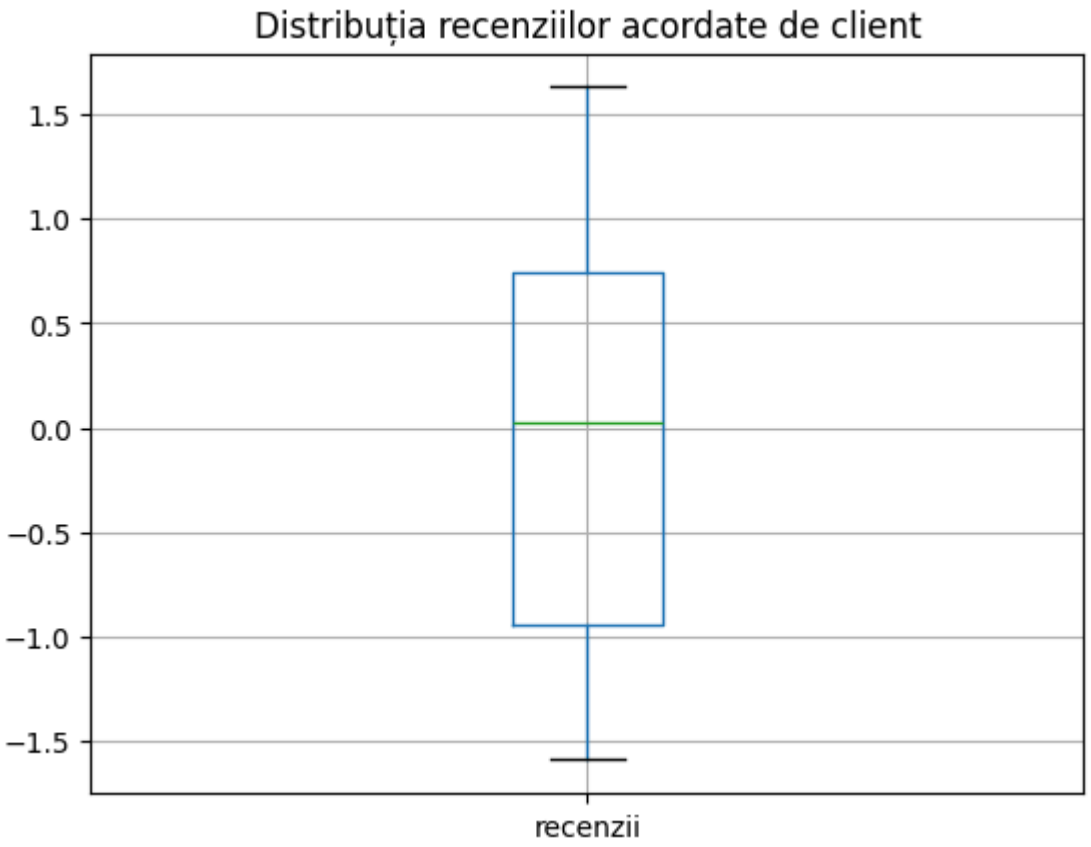

Distribuția varstelor clientilor
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
varsta

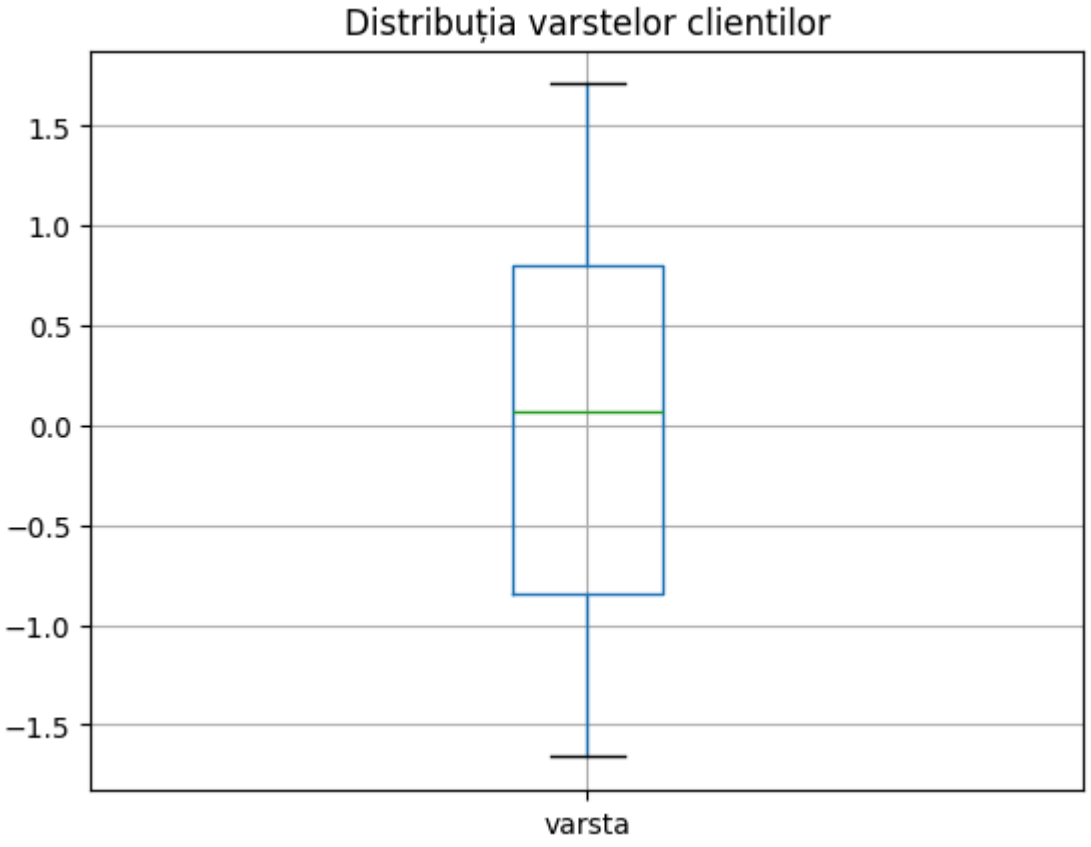

Distribuția valorilor medii ale comenzilor
2
1
0
−1
−2
valoare_comanda_medie


Distribuția recenziilor acordate de client
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
recenzii

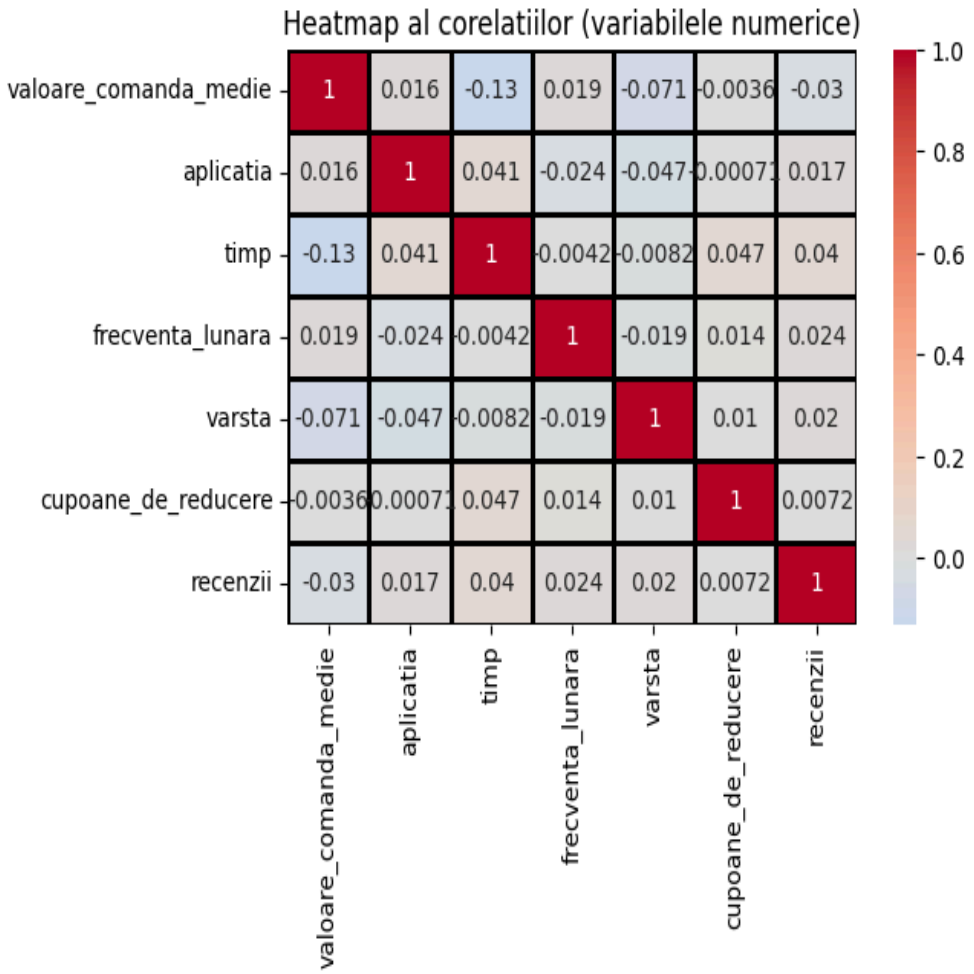

Distribuția frecventei lunare pe magazin
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
frecventa_lunara

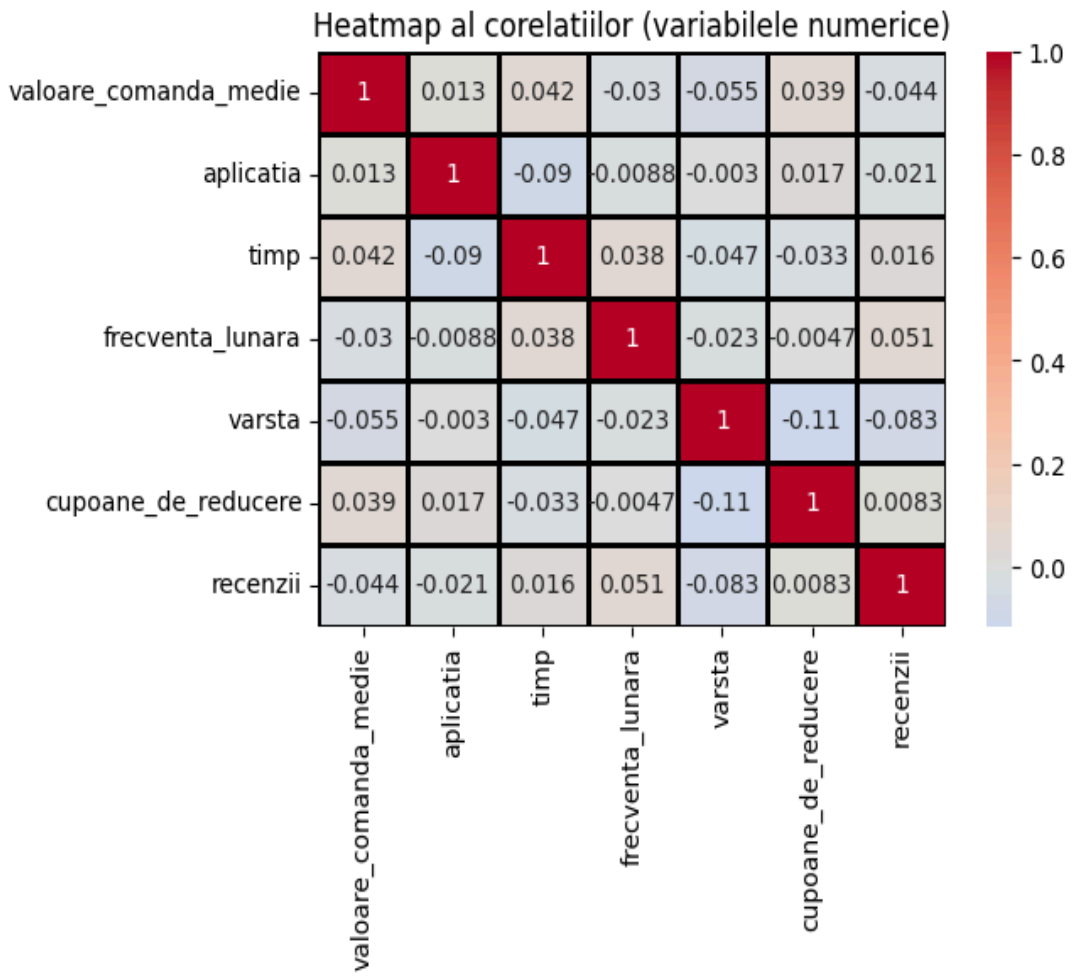

Distribuția timpului petrecut pe magazin
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
timp



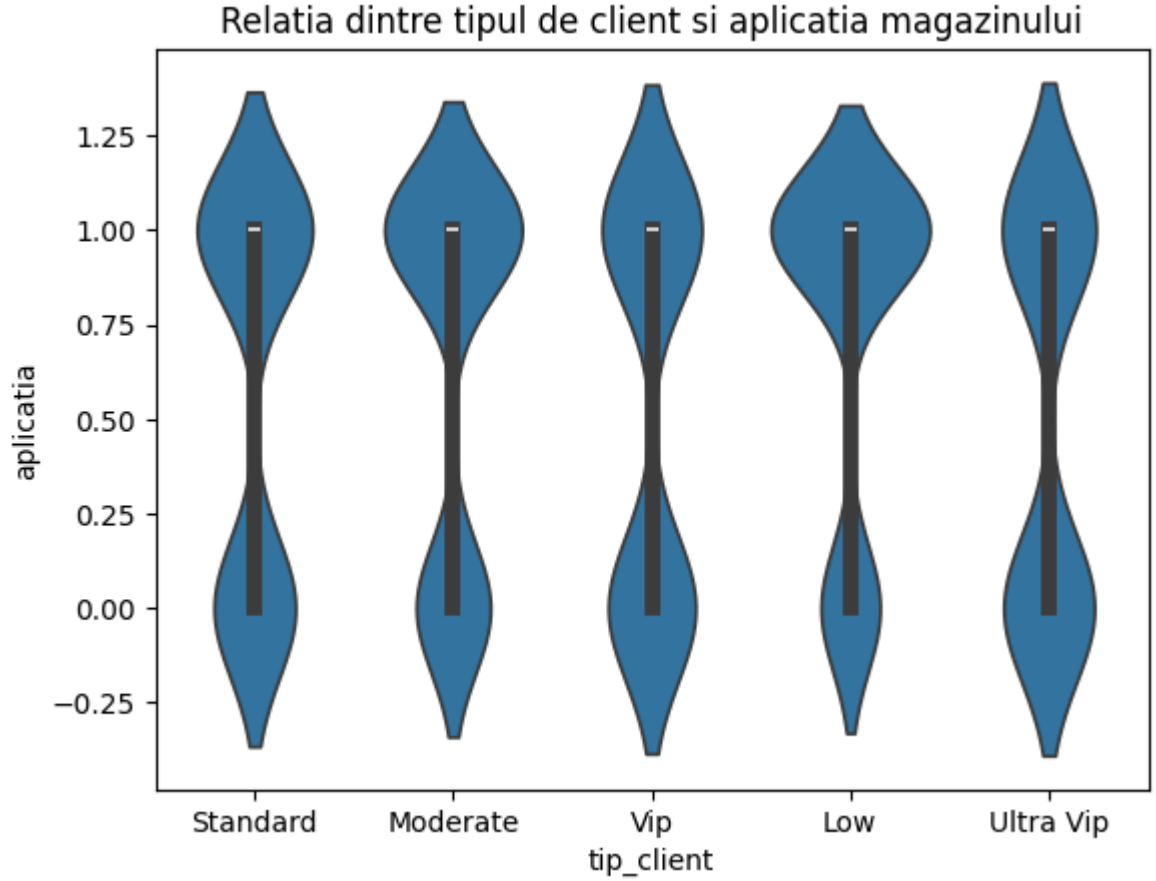


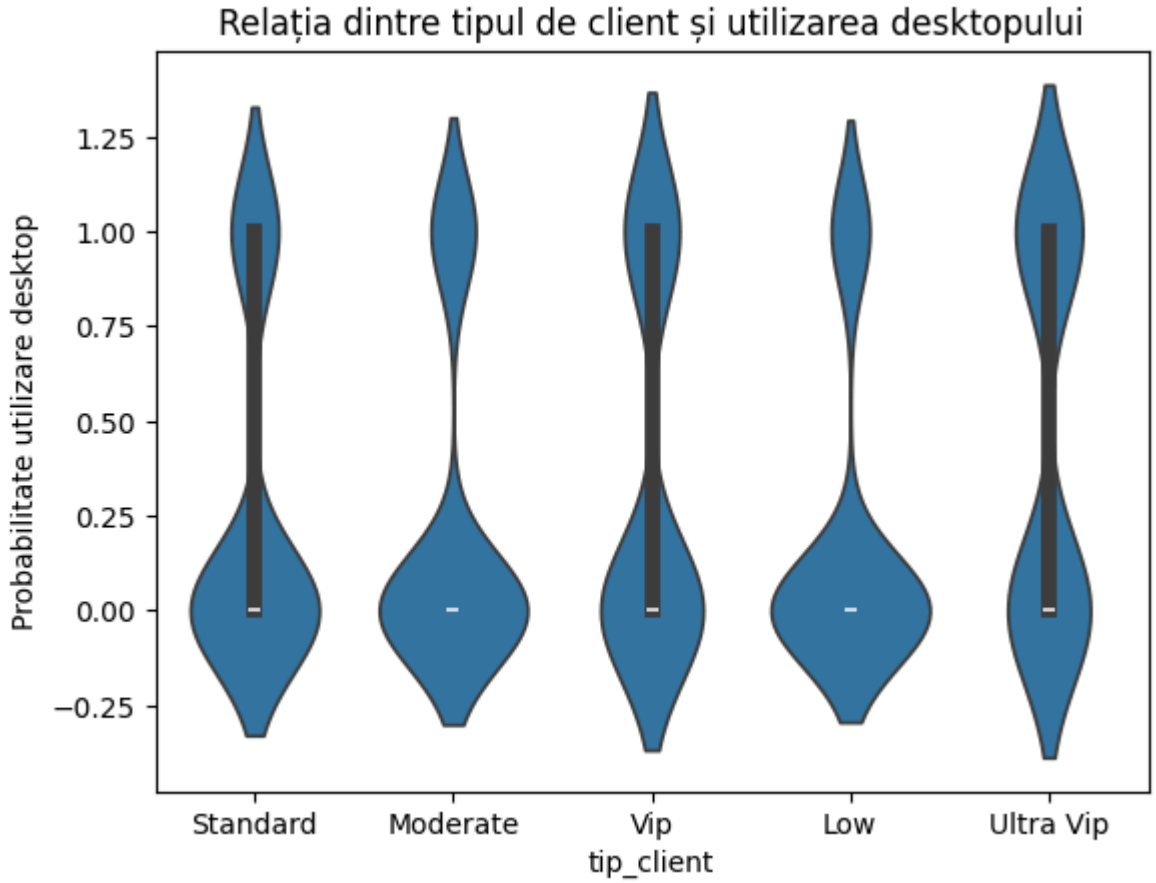
2 .e) Analiza corelatiilor: Am generat un heatmap cu coeficientii de corelatie.

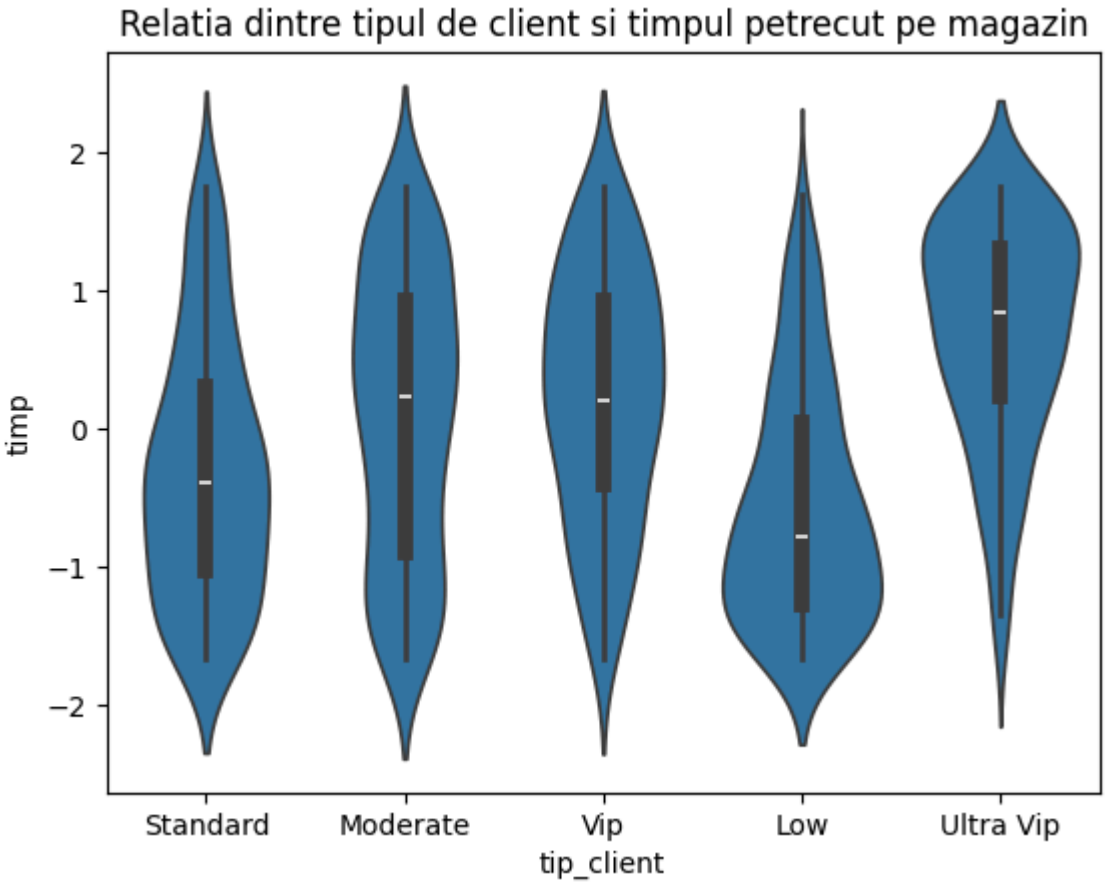

Heatmap al corelatiilor (variabilele numerice)
valoare_comanda_medie
aplicatia
timp
frecventa_lunara
varsta
cupoane_de_reducere
recenzii
1 0.016 -0.13 0.019 -0.071 -0.0036 -0.03
0.016 1 0.041 -0.024 -0.047 0.00071 0.017
-0.13 0.041 1 -0.0042 -0.0082 0.047 0.04
0.019 -0.024 -0.0042 1 -0.019 0.014 0.024
-0.071 -0.047 -0.0082 -0.019 1 0.01 0.02
-0.0036 0.00071 0.047 0.014 0.01 1 0.0072
-0.03 0.017 0.04 0.024 0.02 0.0072 1
1.0
0.8
0.6
0.4
0.2
0.0

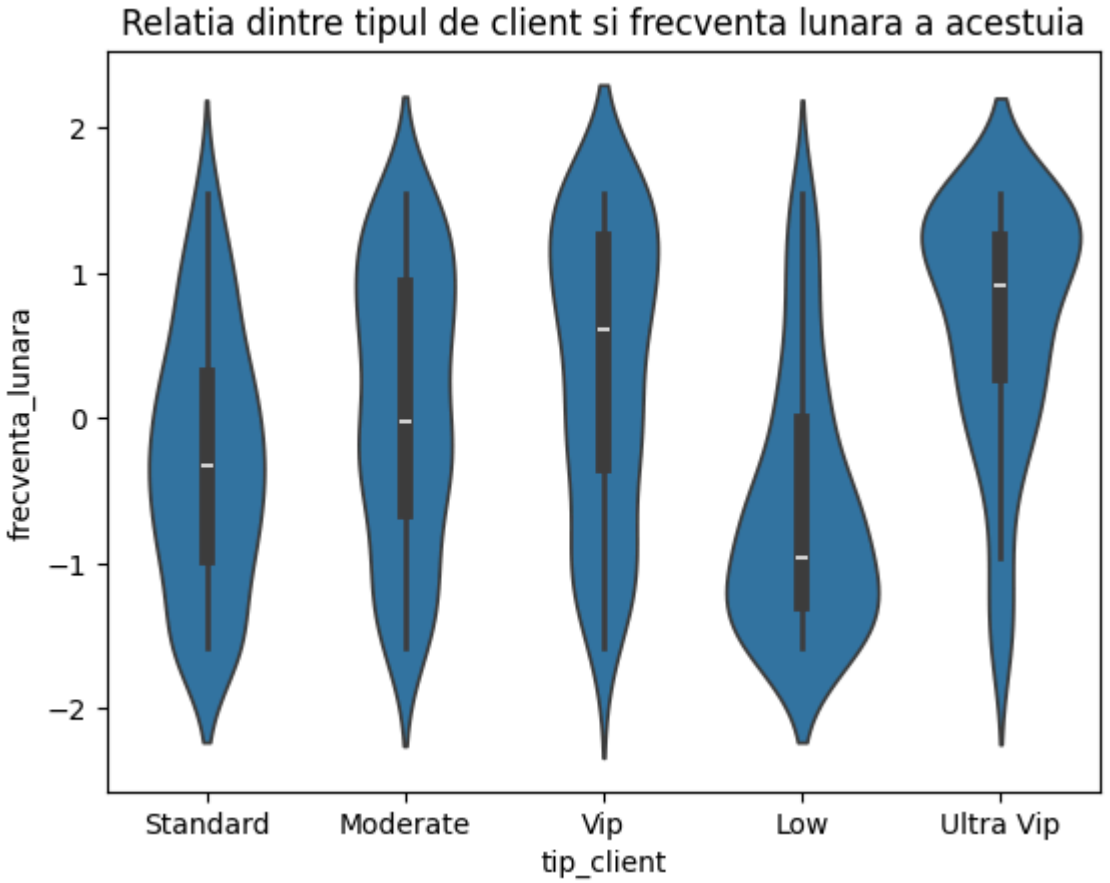


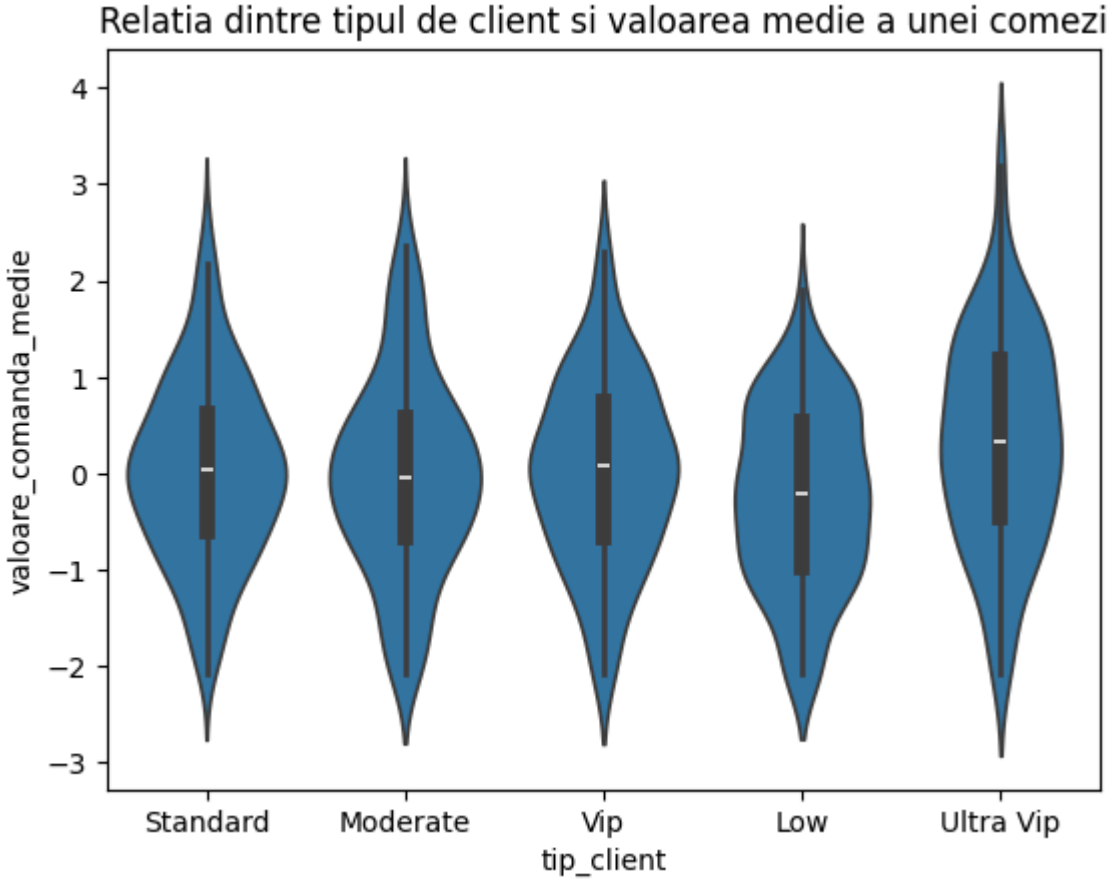
2. f) Relatii intre trasaturi si coloana tinta (tip_client): Pentru a analiza mai profund legatura dintre fiecare trasatura si tip_client, am folosit Violin plots (pentru a vedea distributiile in functie de clasa) si am inclus atat datele de train, cat si cele de test pentru comparatie (relatia dintre tip_client si utilizarea aplicatiei, relatia dintre tip_client si varsta/frecventa comenzilor, impactul metodei de plata sau a cupoanelor de rudecere).

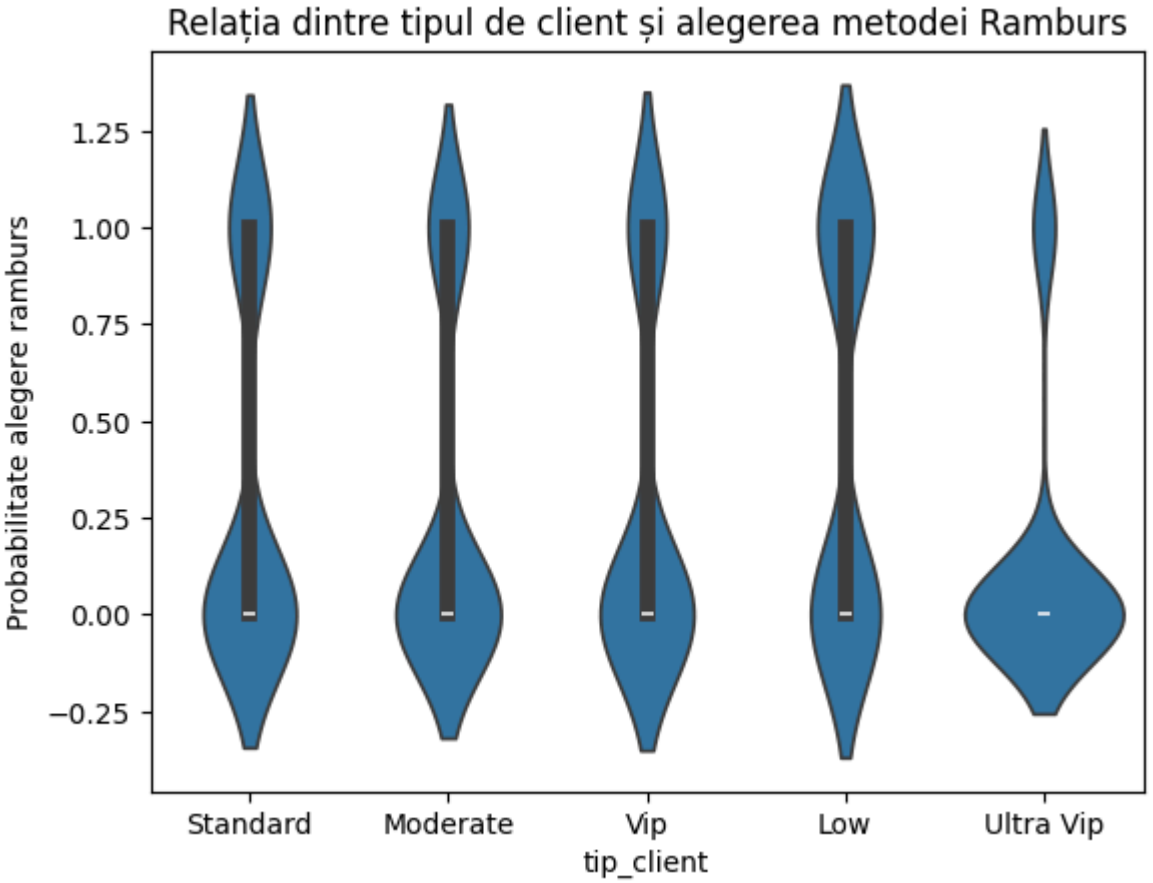

Relatia dintre tipul de client si aplicatia magazinului
aplicatia
1.25
1.00
0.75
0.50
0.25
0.00
−0.25
Standard
Moderate
Vip
Low
Ultra Vip
tip_client

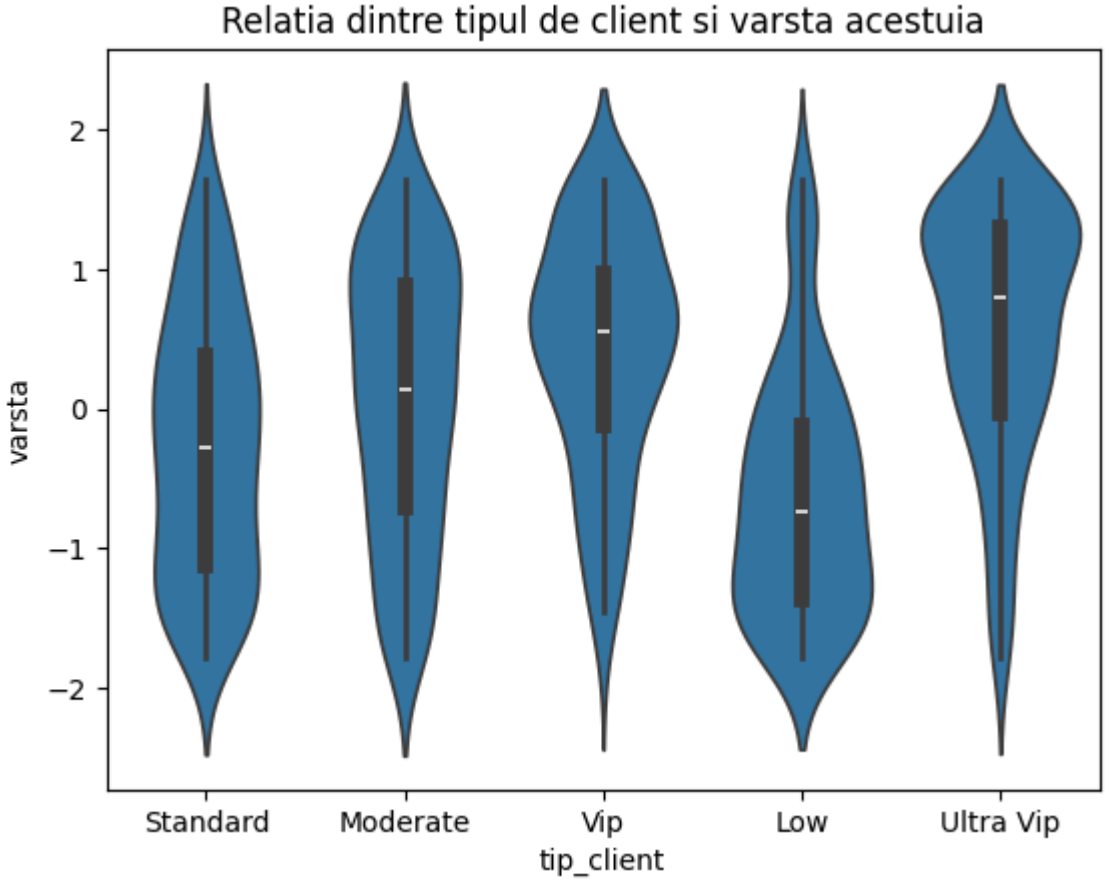

Relația dintre tipul de client și utilizarea desktopului
Probabilitate utilizare desktop
1.25
1.00
0.75
0.50
0.25
0.00
−0.25
Standard
Moderate
Vip
Low
Ultra Vip
tip_client

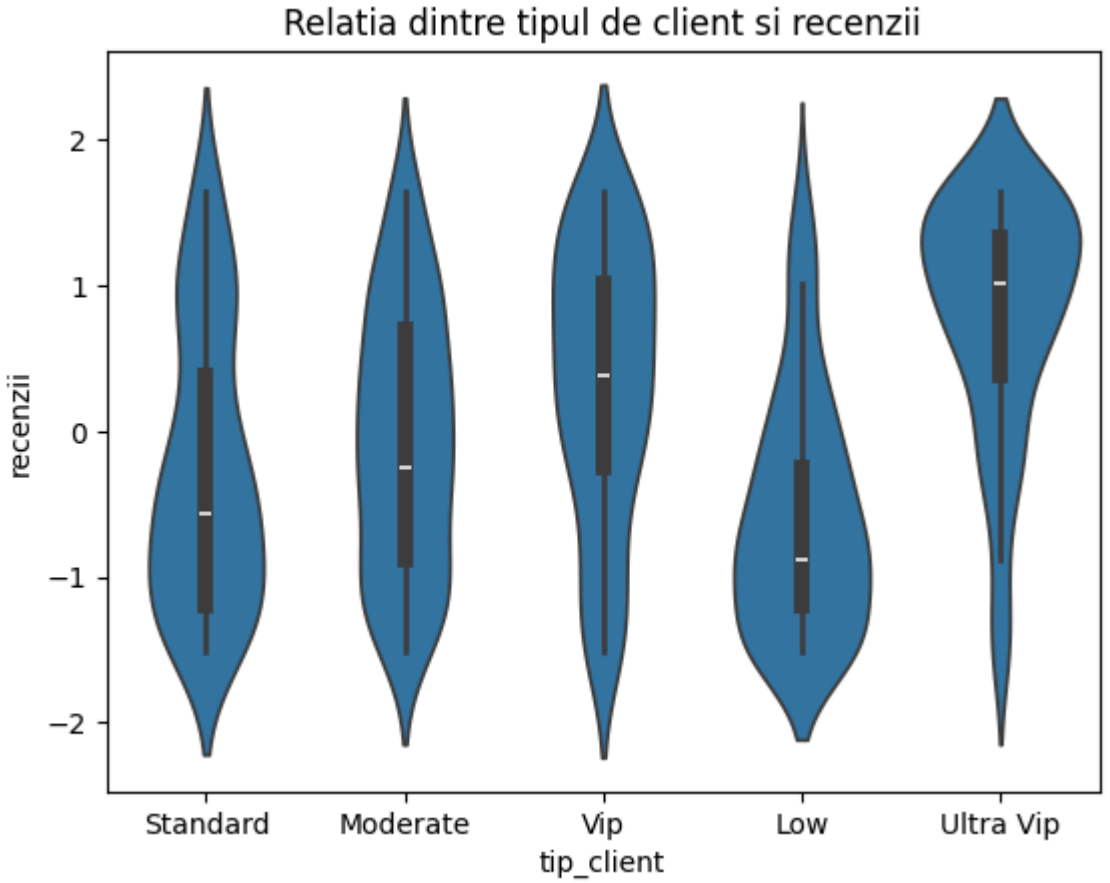

Relatia dintre tipul de client si timpul petrecut pe magazin
timp
2
1
0
−1
−2
Standard
Moderate
Vip
Low
Ultra Vip
tip_client

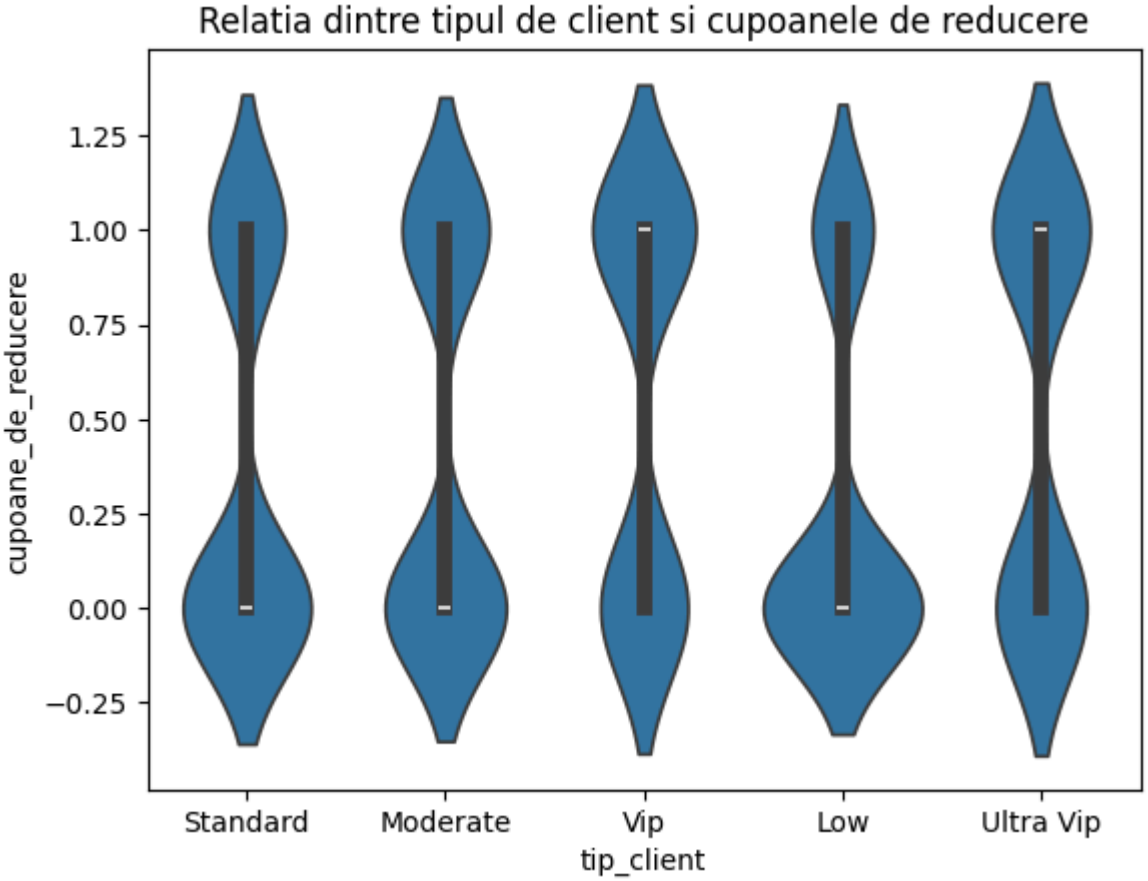

Relatia dintre tipul de client si frecventa lunara a acestuia
frecventa_lunara
2
1
0
−1
−2
Standard
Moderate
Vip
Low
Ultra Vip
tip_client

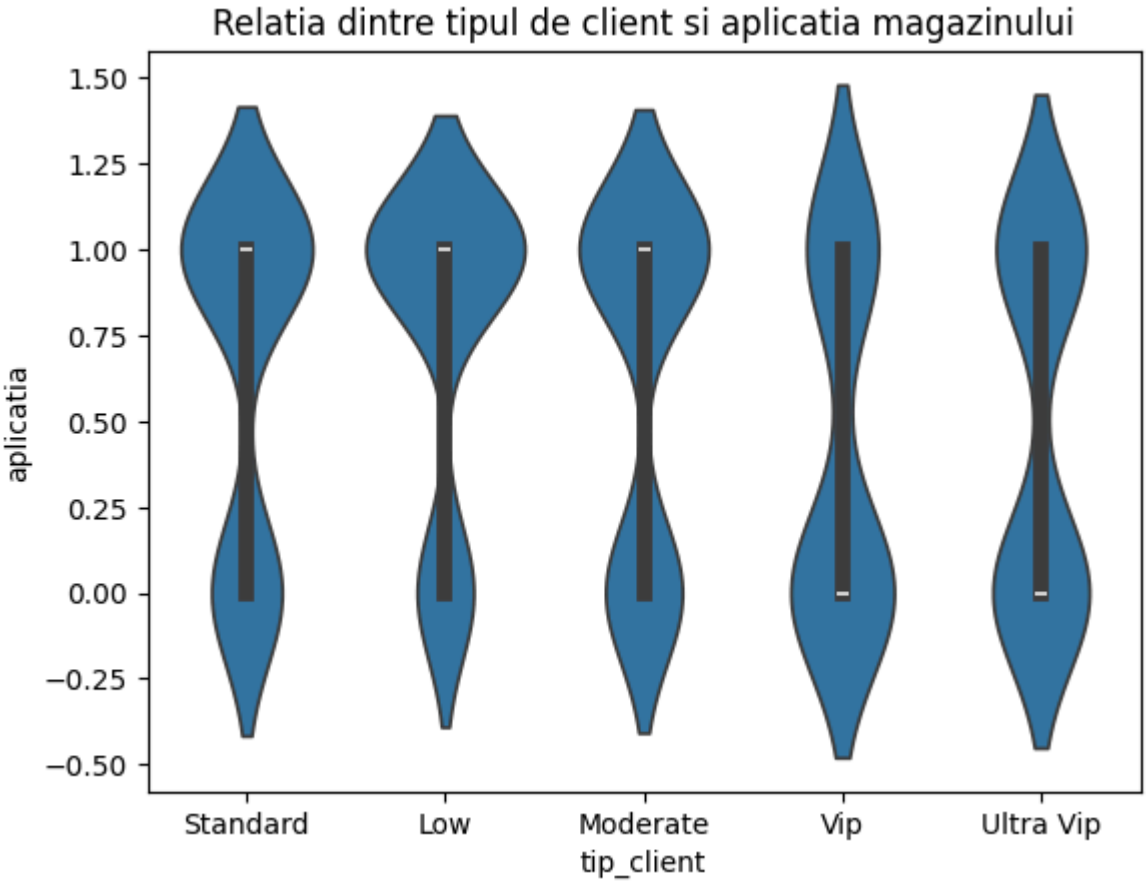

Relatia dintre tipul de client si valoarea medie a unei comezi
valoare_comanda_medie
4
3
2
1
0
−1
−2
−3
Standard
Moderate
Vip
Low
Ultra Vip
tip_client

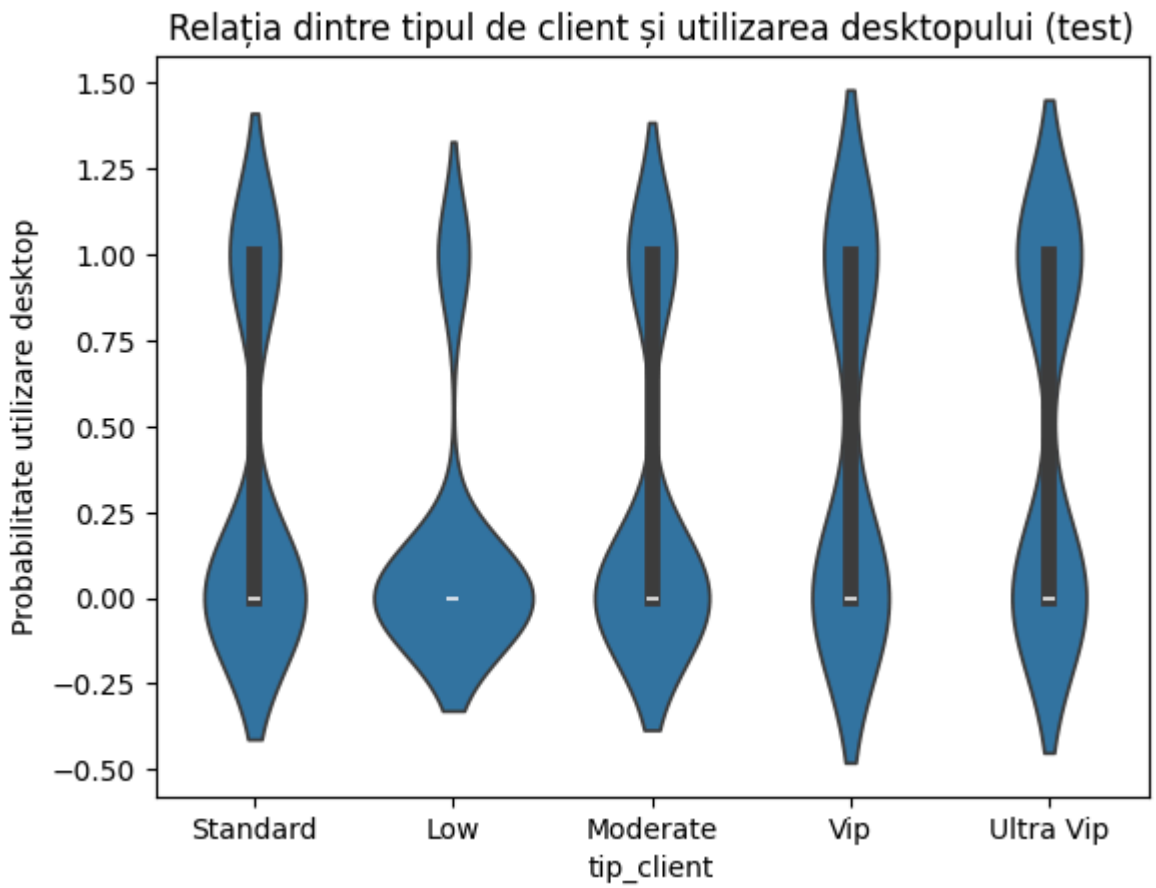

Relația dintre tipul de client și alegerea metodei Ramburs
Probabilitate alegere ramburs
1.25
1.00
0.75
0.50
0.25
0.00
−0.25
Standard
Moderate
Vip
Low
Ultra Vip
tip_client

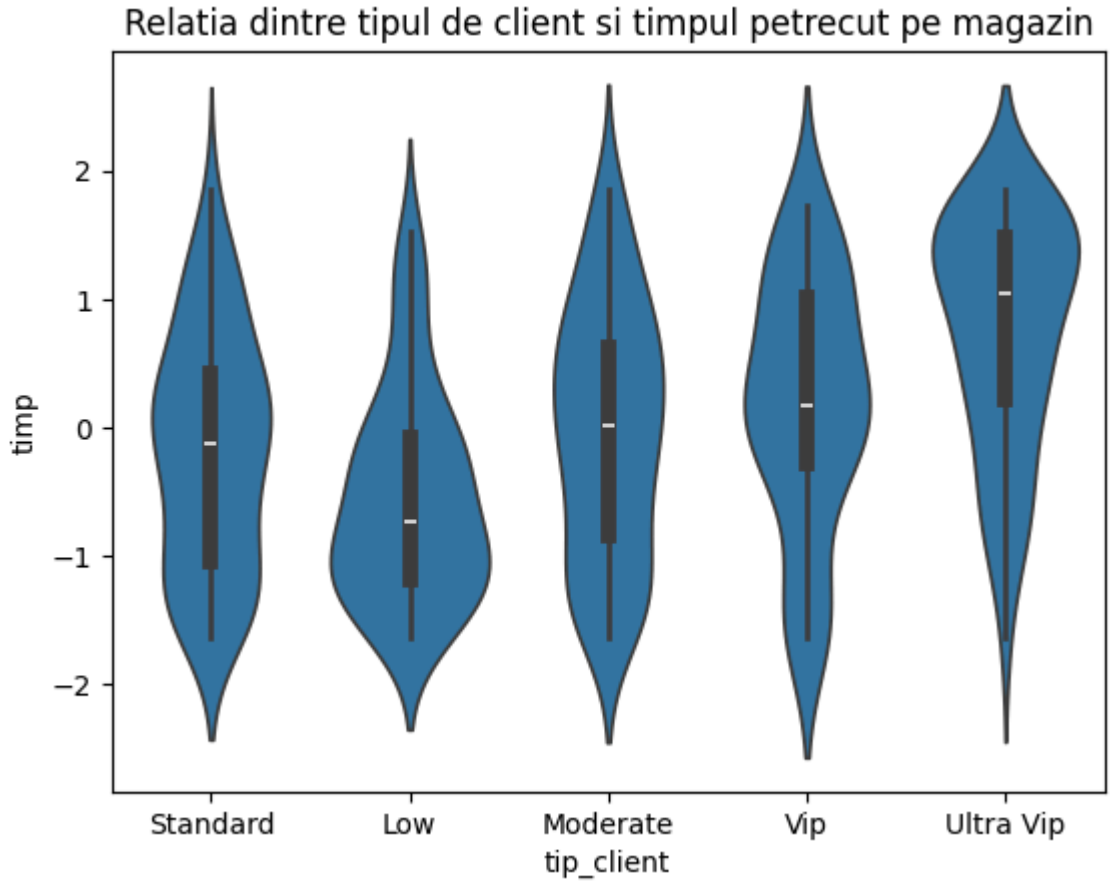

Relatia dintre tipul de client si varsta acestuia
varsta
2
1
0
−1
−2
Standard
Moderate
Vip
Low
Ultra Vip
tip_client

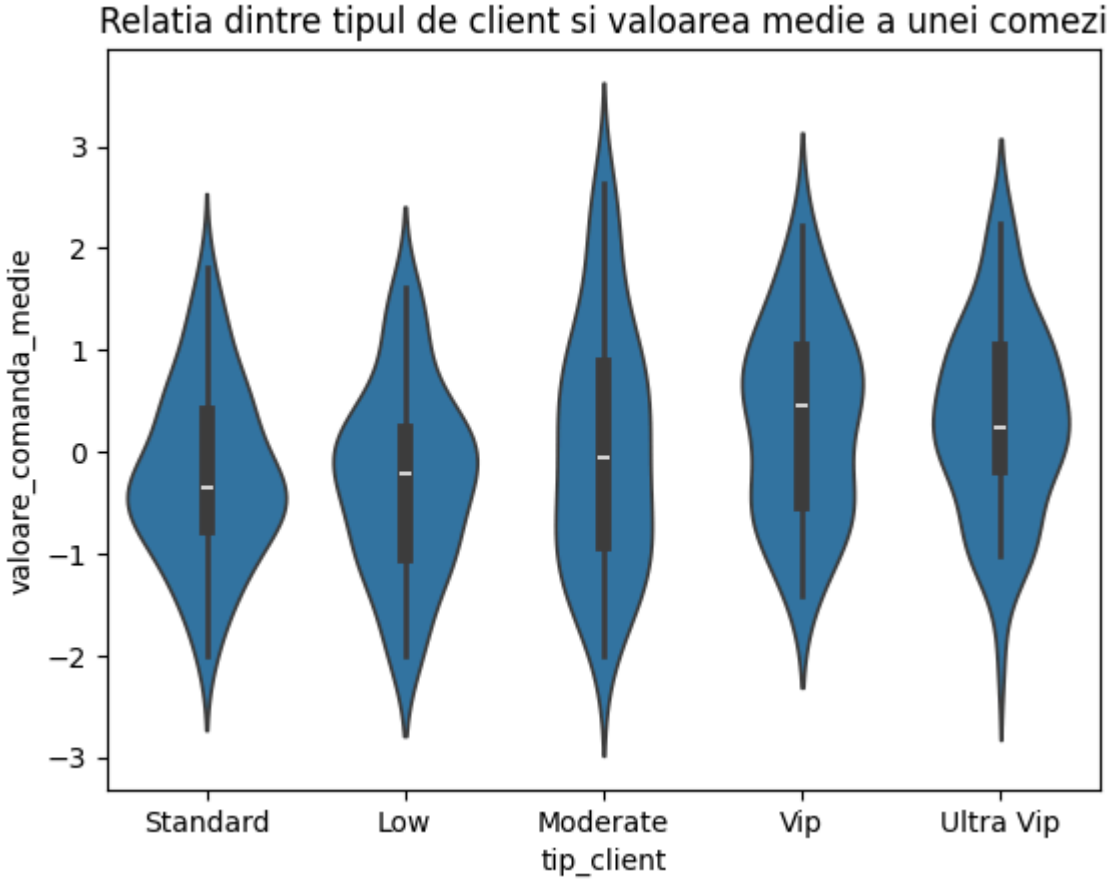

Relatia dintre tipul de client si recenzii
recenzii
2
1
0
−1
−2
Standard
Moderate
Vip
Low
Ultra Vip
tip_client

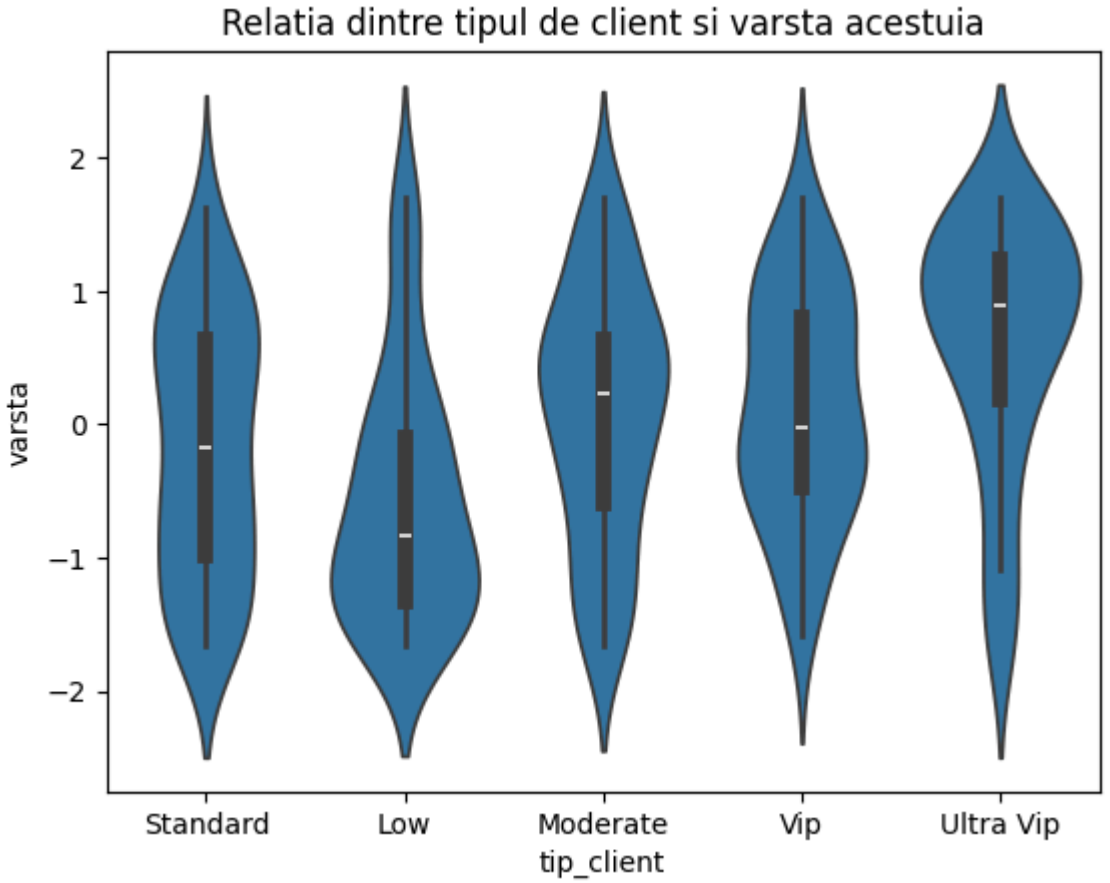

Relatia dintre tipul de client si cupoanele de reducere
cupoane_de_reducere
1.25
1.00
0.75
0.50
0.25
0.00
−0.25
Standard
Moderate
Vip
Low
Ultra Vip
tip_client

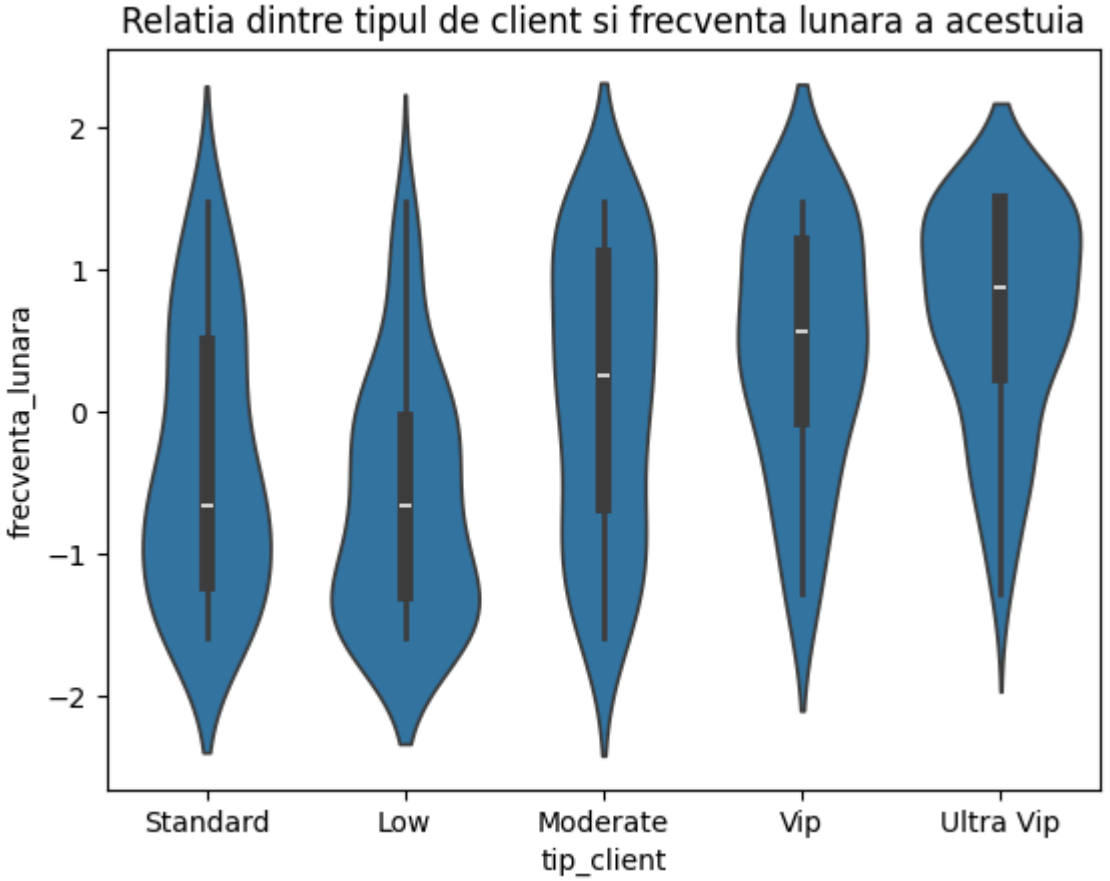

Relatia dintre tipul de client si aplicatia magazinului
aplicatia
1.50
1.25
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
Standard
Low
Moderate
Vip
Ultra Vip
tip_client

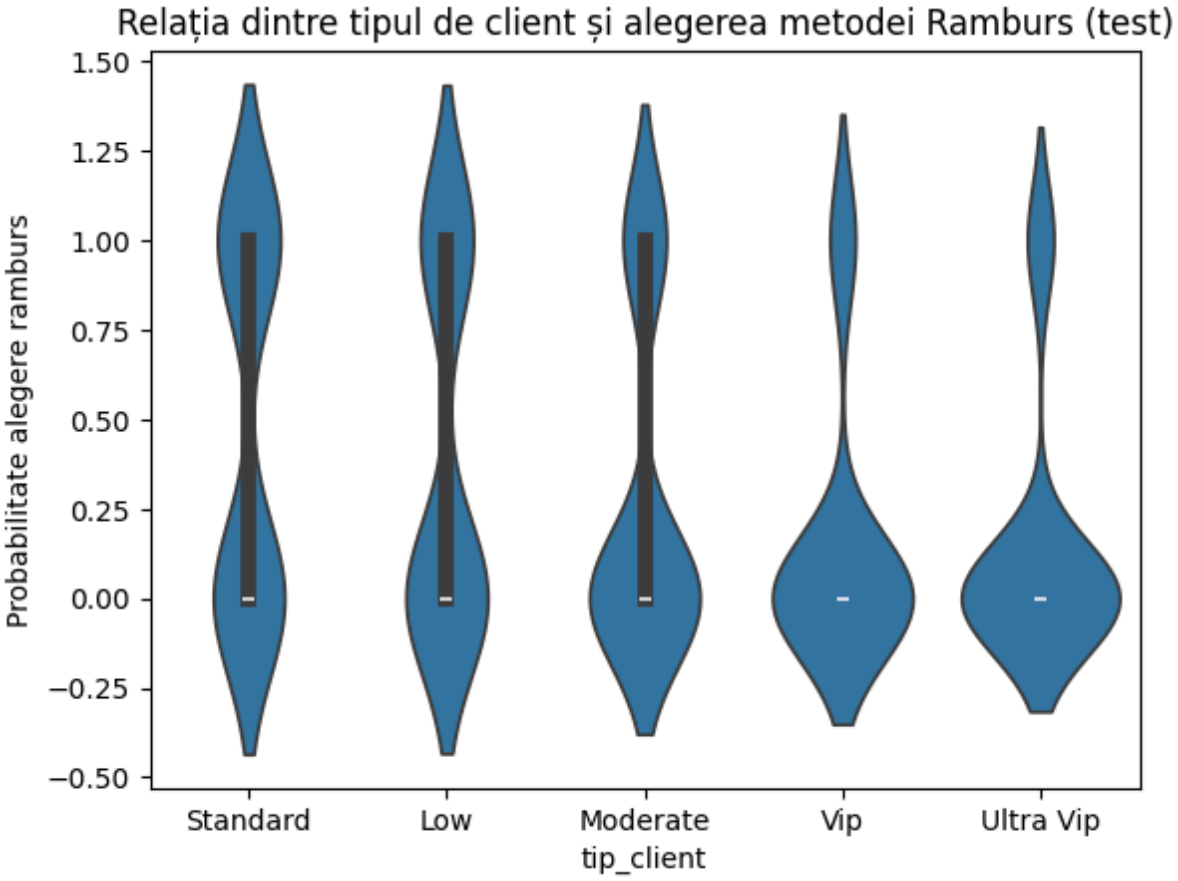

Relația dintre tipul de client și utilizarea desktopului (test)
Probabilitate utilizare desktop
1.50
1.25
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
Standard
Low
Moderate
Vip
Ultra Vip
tip_client

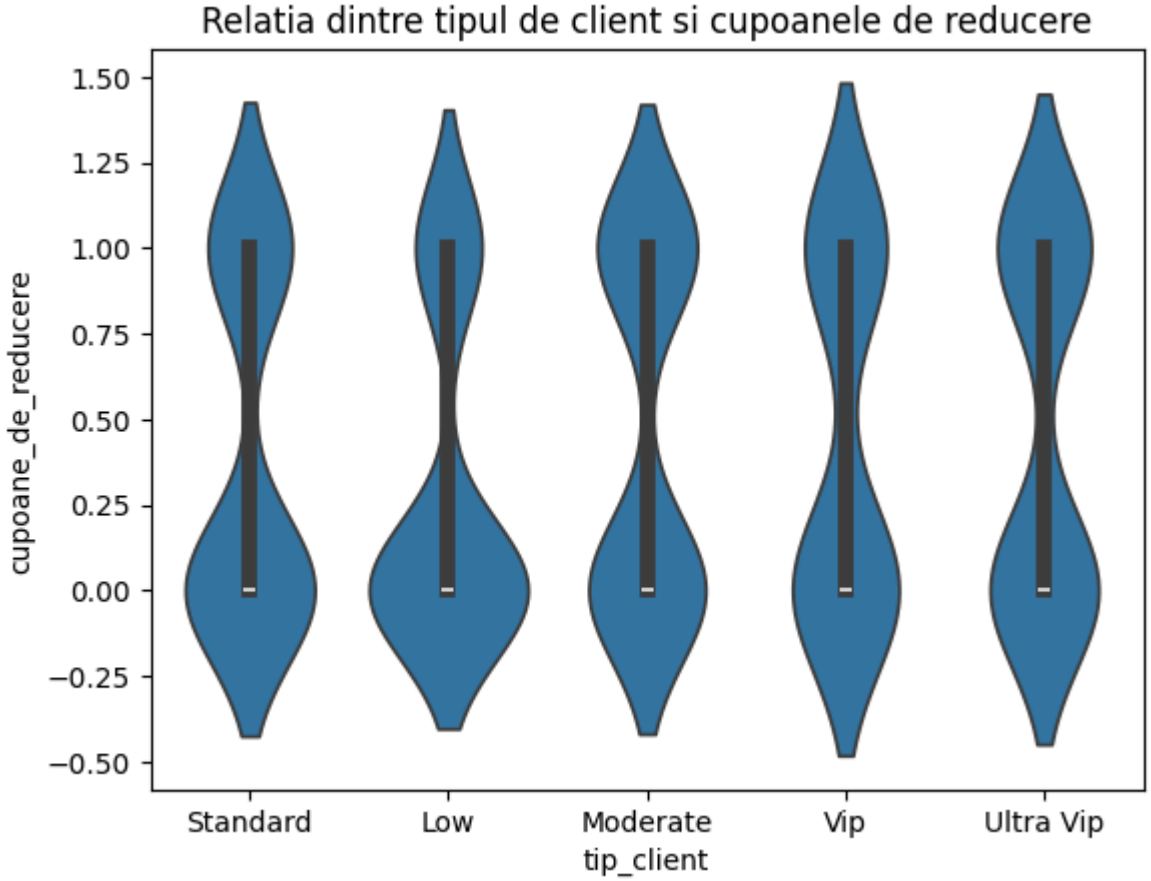

Relatia dintre tipul de client si timpul petrecut pe magazin
timp
2
1
0
−1
−2
Standard
Low
Moderate
Vip
Ultra Vip
tip_client

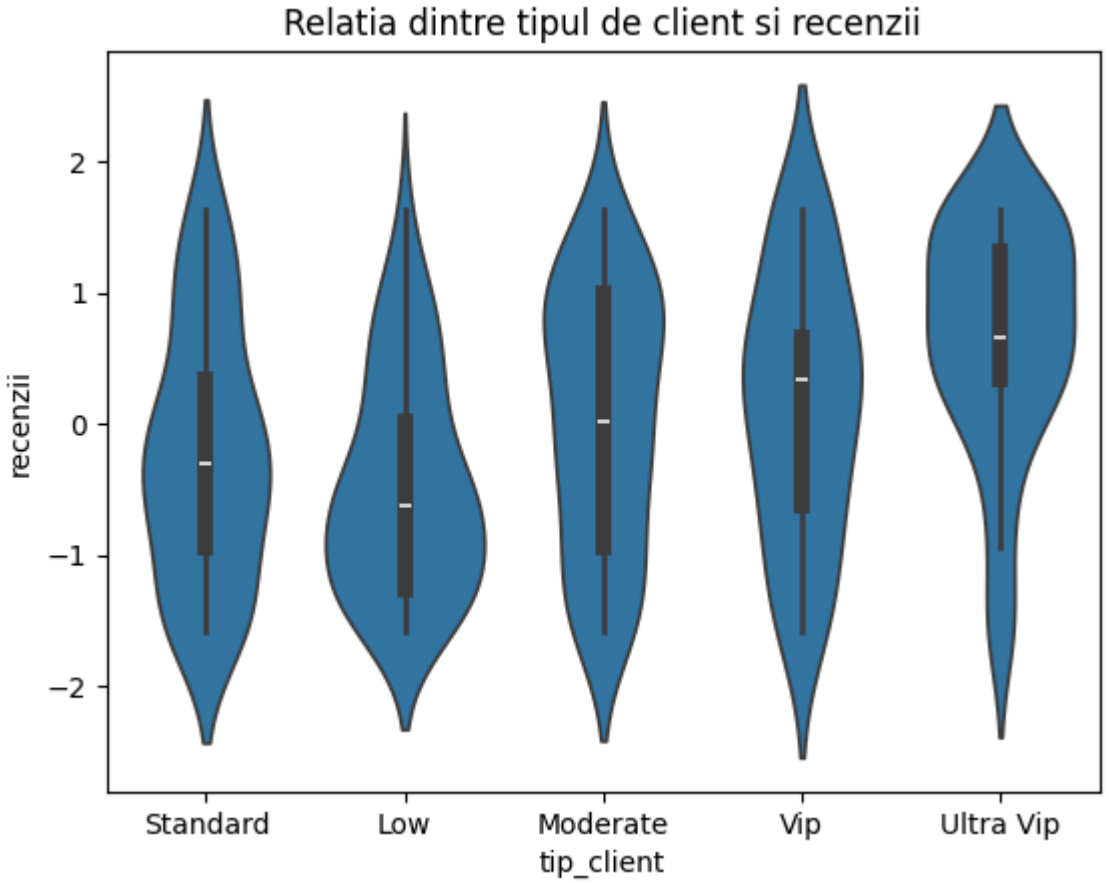

Relatia dintre tipul de client si valoarea medie a unei comezi
valoare_comanda_medie
3
2
1
0
−1
−2
−3
Standard
Low
Moderate
Vip
Ultra Vip
tip_client


Relatia dintre tipul de client si varsta acestuia
varsta
2
1
0
−1
−2
Standard
Low
Moderate
Vip
Ultra Vip
tip_client


Relatia dintre tipul de client si frecventa lunara a acestuia
frecventa_lunara
2
1
0
−1
−2
Standard
Low
Moderate
Vip
Ultra Vip
tip_client


Relația dintre tipul de client și alegerea metodei Ramburs (test)
Probabilitate alegere ramburs
1.50
1.25
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
Standard
Low
Moderate
Vip
Ultra Vip
tip_client


Relatia dintre tipul de client si cupoanele de reducere
cupoane_de_reducere
1.50
1.25
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
Standard
Low
Moderate
Vip
Ultra Vip
tip_client


Relatia dintre tipul de client si recenzii
recenzii
2
1
0
−1
−2
Standard
Low
Moderate
Vip
Ultra Vip
tip_client

3. Antrenarea si compararea a 3 algoritmi diferiti si evaluarea performantei

3. a) Impartirea datelor in antrenament si test: Setul de date train_encodat.csv a fost impartit in X_train (variabilele independente fara coloana tip_client) si y_train (variabila dependenta tip_client). Setul de date test_encodat.csv a fost pregatit in acelasi mod pentru testare (X_test, y_test).

3. b) Alegerea si antrenarea modelelor: Am ales trei modele clasice: Regresie Logistica, Random Forest, KNN. Toate cele trei modele au fost antrenate folosind datele din X_train si y_train.

3. c) Precizarea rezultatelor: Fiecare model a fost folosit pentru a face predictii asupra setului de test X_test. Primele 10 predictii au fost comparate vizual cu valorile reale din y_test.

3. d) Evaluarea performantei: Au fost calculate acuratetea (cat de multe clasificari au fost corecte) si scorul F1 (media armonica intre precizie si recall)

3. e) Matricea de confuzie: Pentru fiecare model am generat o matrice de confuzie care arata cate instante din fiecare clasa au fost corect sau fresit clasificate si distributia erorilor.

3. f) Tabel comparativ de scoruri: Un tabel final comparativ a fost creat cu: numele fiecarui model, acuratetea si scorul F1 corespunzator.