Constantinescu Rares - 313CA

Se realizeaza implementarea unui model care verifica tranzactii bancare frauduloase.

Tipul problemei este de clasificare.

Setul de date principal este generat de generate_data.py, care creeaza un set de data in format csv. Setul de date generate are 1000 de linii, a cate 10
parametrii relevanti. Acesti 10 parametrii au diverse tipuri de date, spre exemplu numere intregi, reale sau stringuri.

Setul de date principal (all_data.csv) este impartit in doua seturi de date de catre split_data.csv. Dupa impartire, test.csv va avea 25% din exemplele din all_data.csv, iar train.csv va avea 75% din exemplele din all_data.csv.

Setul de date de testare este folosit pentru a evalua acuratetea modelului, iar setul de date de antrenare este folosit pentru a antrena modelul.

Setul de date genereaza date random pentru diversi parametrii, folosind seedul 12345. Data si ora generate sunt random, din ultimul an. Locatiile, dispozitivele
si tipurile de tranzactii sunt generate random, dintr-o lista data de mana pentru fiecare dintre ele. Sumele tranzactionate sunt numere intre 1 si 3000 cu 2 zecimale.
Probabilitatea ca o tranzactie sa fie de tip fraud se foloseste de un bias care presupuna ca tranzactiile unor sume mai mari sunt mai probabile sa fie frauduloase,
asa ca daca suma tranzactiei este mai mare de 2250, probabilitatea ca aceasta sa fie frauduloasa este de 85%, altfel este de 3%. Fraudalitatea este 0/1 in
functie de probabilitatea de fraudare calculata random anterior.

EDA.py contine implementarea principala a cerintelor.

-6a) este realizat prin dropna

-6b) scrie in fisierele _descriptive rezultatele obtinute pentru coloanele numerice, respectiv caegorice

-6c) pentru variabile numerice salveaza histogramele, iar pentru cele categorice gaseste top 10 cele mai folosite categorii si salveaza bar plotul acestora

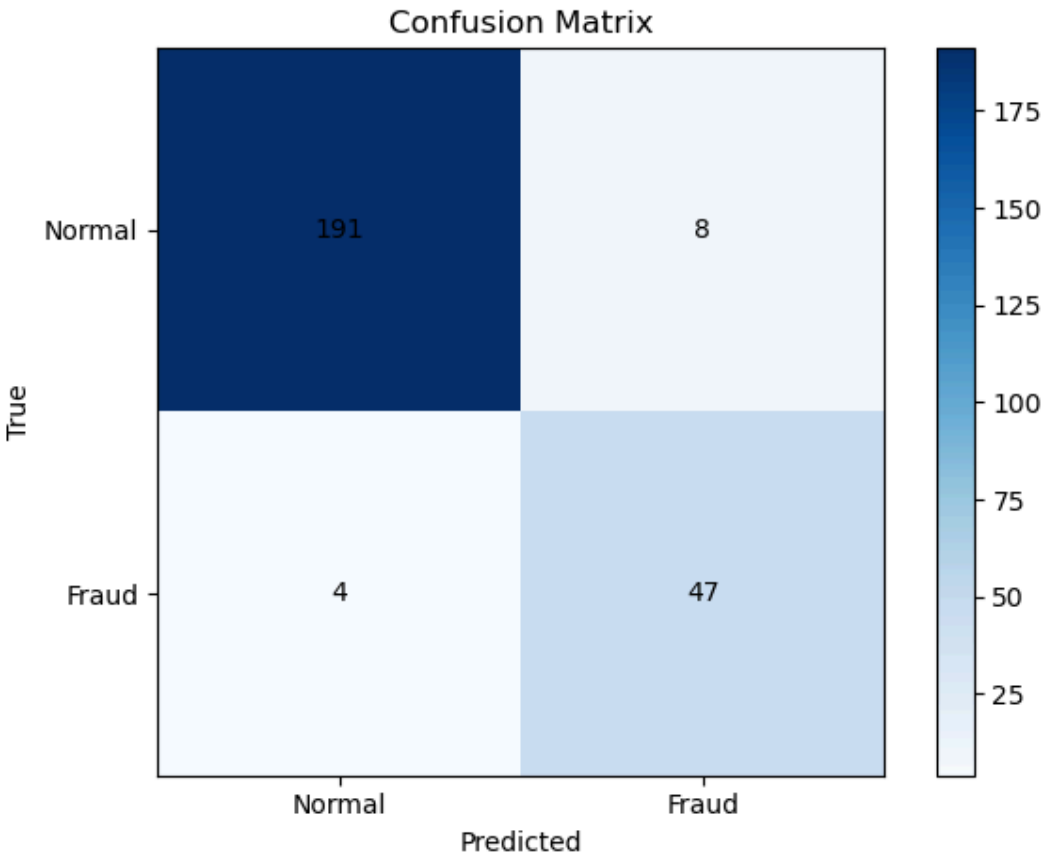
-6d) genereaza box ploturi pentru a identifica valorile aberante

-6e) calculeaza corelatia dintre coloanele numerice si o afiseaza ca un heatmap

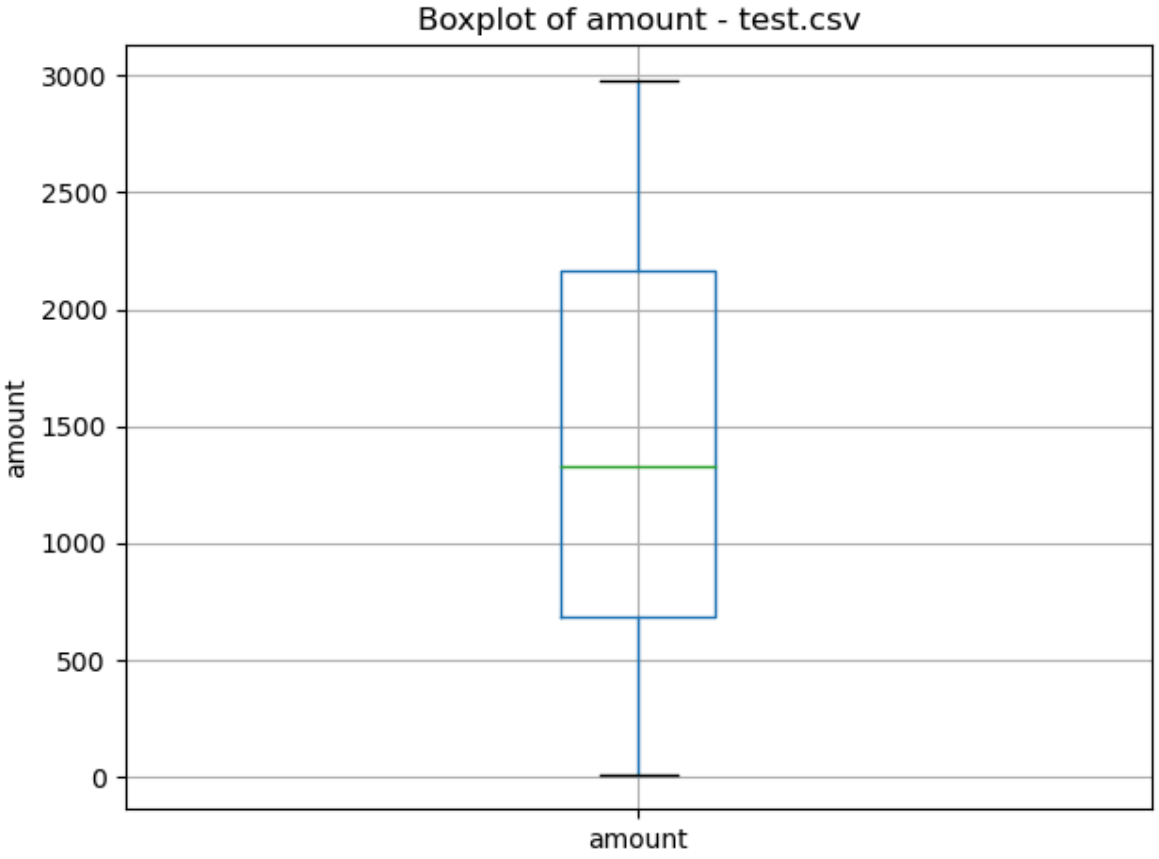
-6f) se foloseste de violin ploturi pentru a arata distributia valorilor unei caracteristici pentru fiecare clasa

-7 Elimina liniile cu date lipsa. Transforma variabilele categorice in valori numerice. Foloseste un model random forest. Evalueaza modelul antrenat pe train.csv
pe setul de test si salveaza: acuratetea, precizia, recall-uri si F1-scoreu, aceste informatii fiind stocate in randomforest.txt, pe setul de date creat, modelul avand o acuratete de 95,2%. Modelul genreaza si matricea de confuzie pe care o salveaza si ca imagine.

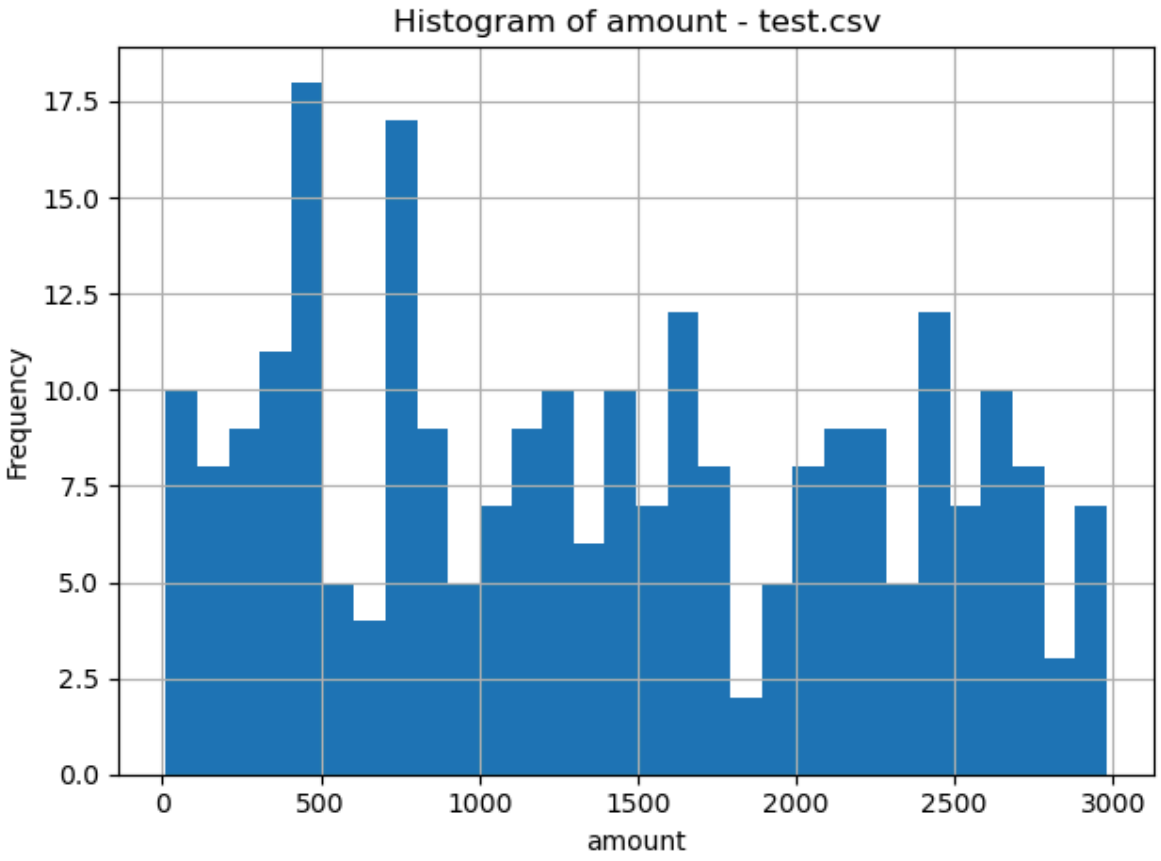
Proiectul contine si o implementare pe github: https://github.com/JKLRares/Bank-Fraud-Analysis



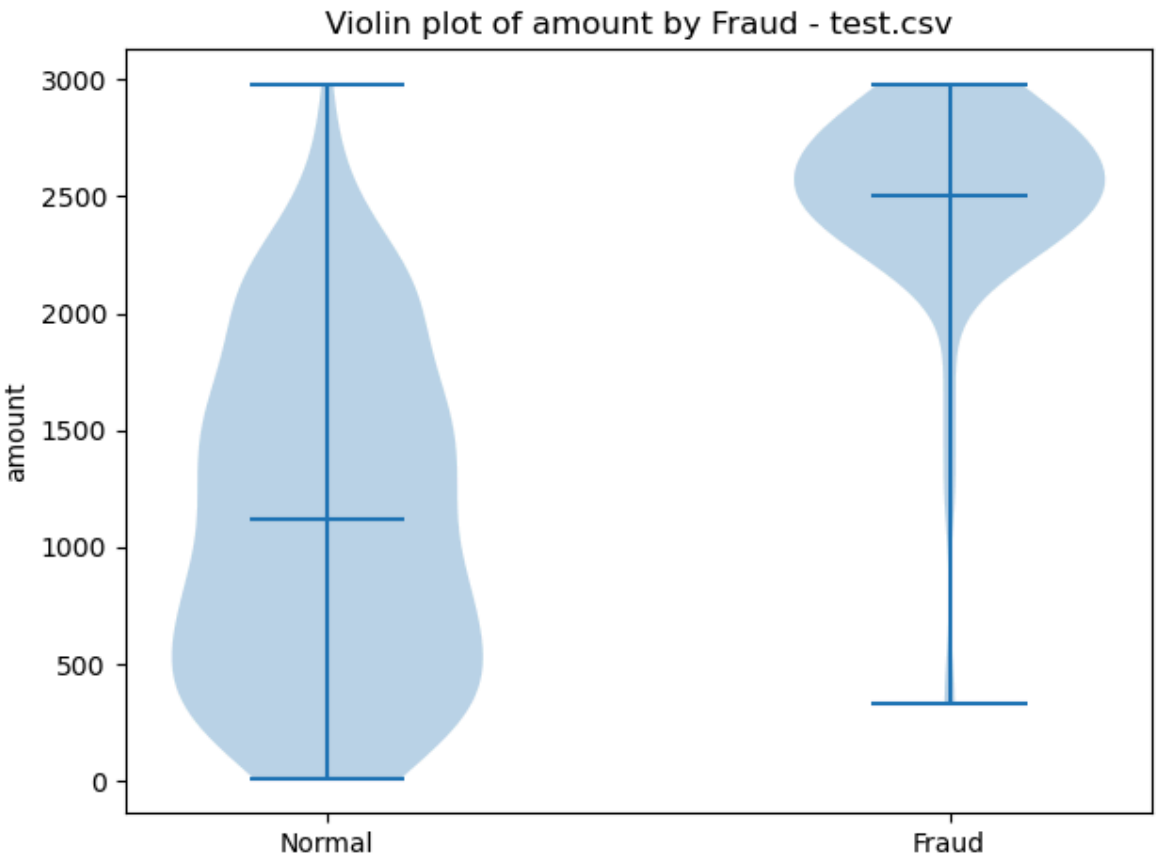
Modelul a clasificat corect 191 de tranzactii normale si 47 frauduloase. Au existat 8 fals pozitive si 4 fals negative



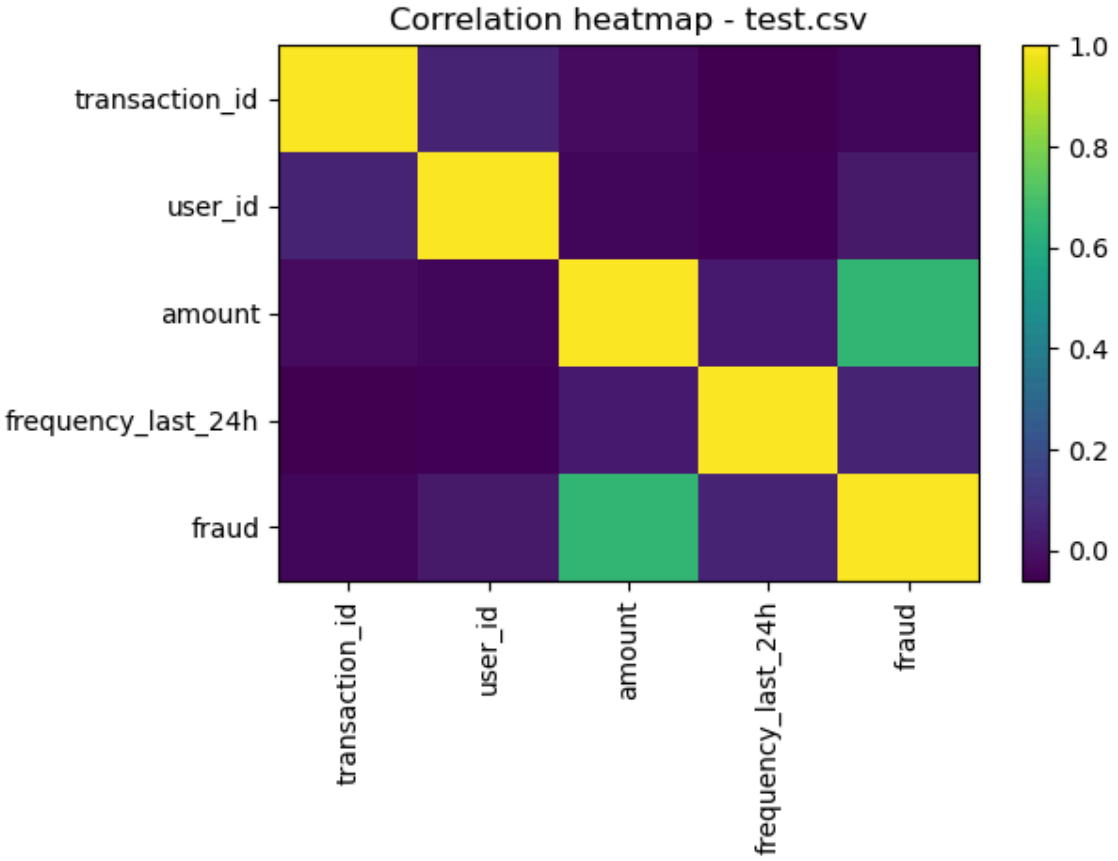
Distributia sumelor



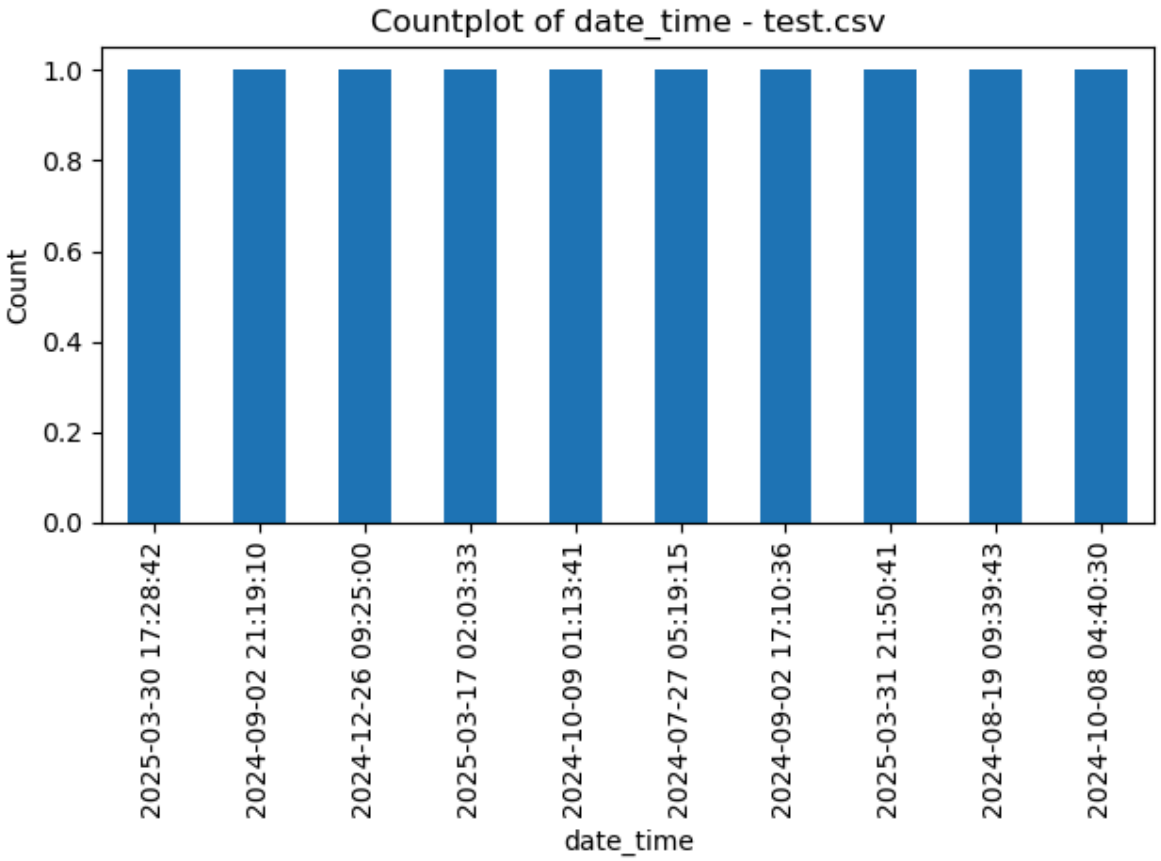
Frecventa sumelor



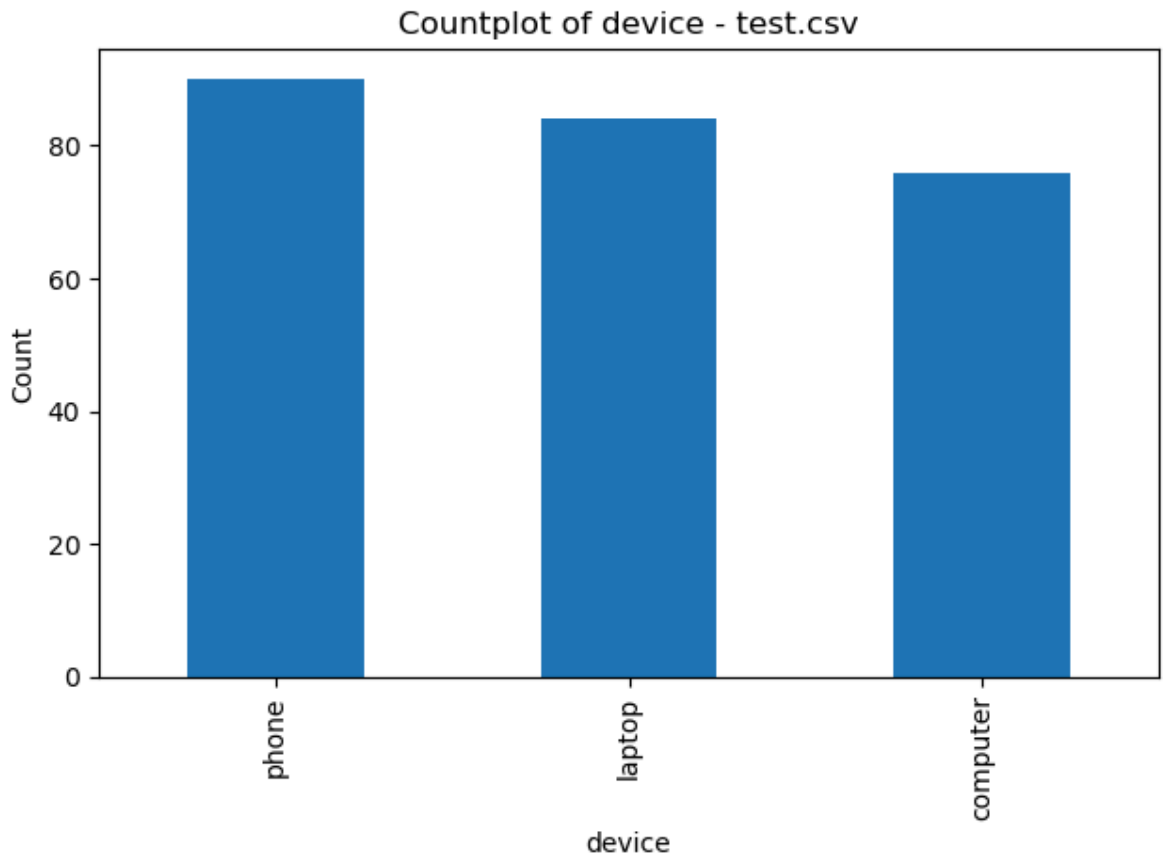
Compara dristributia sumelor pentru tranzactii normale si frauduloase



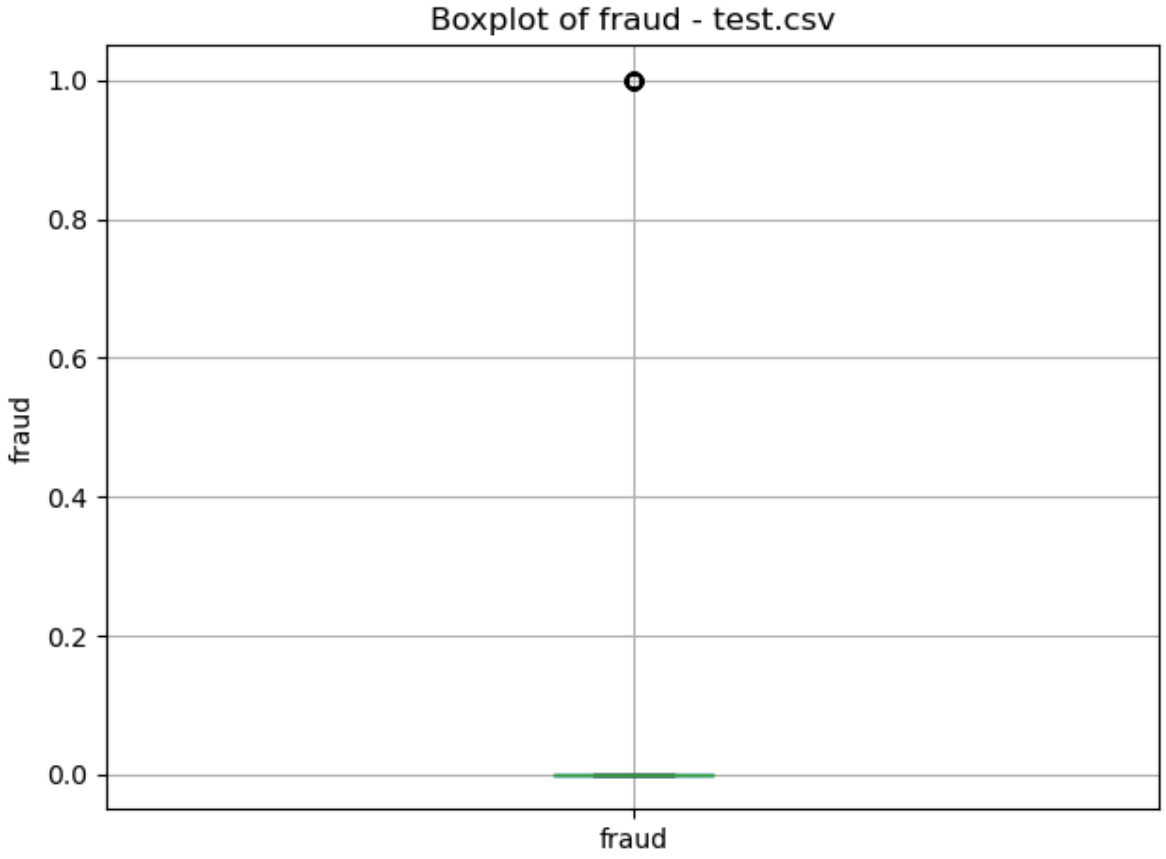
Corelatiile drintre variabilele numerice



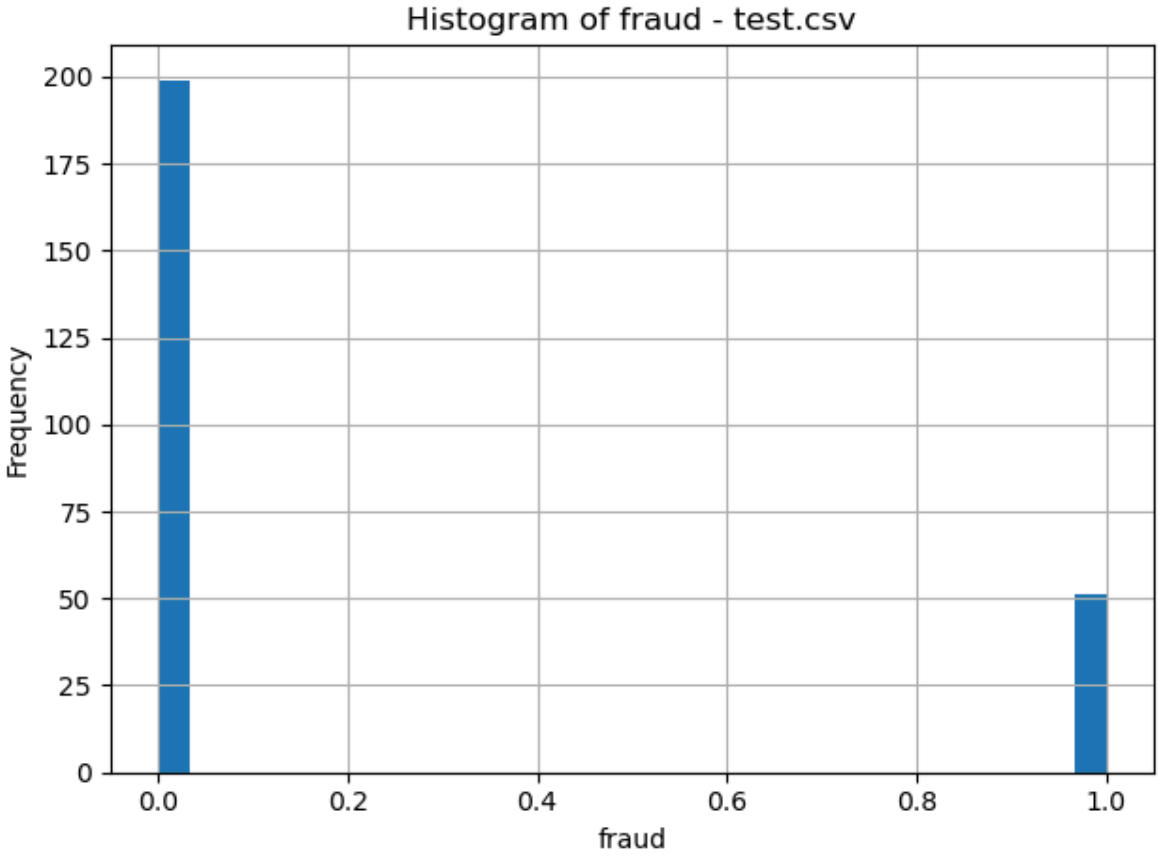
Fiecare data din top 10



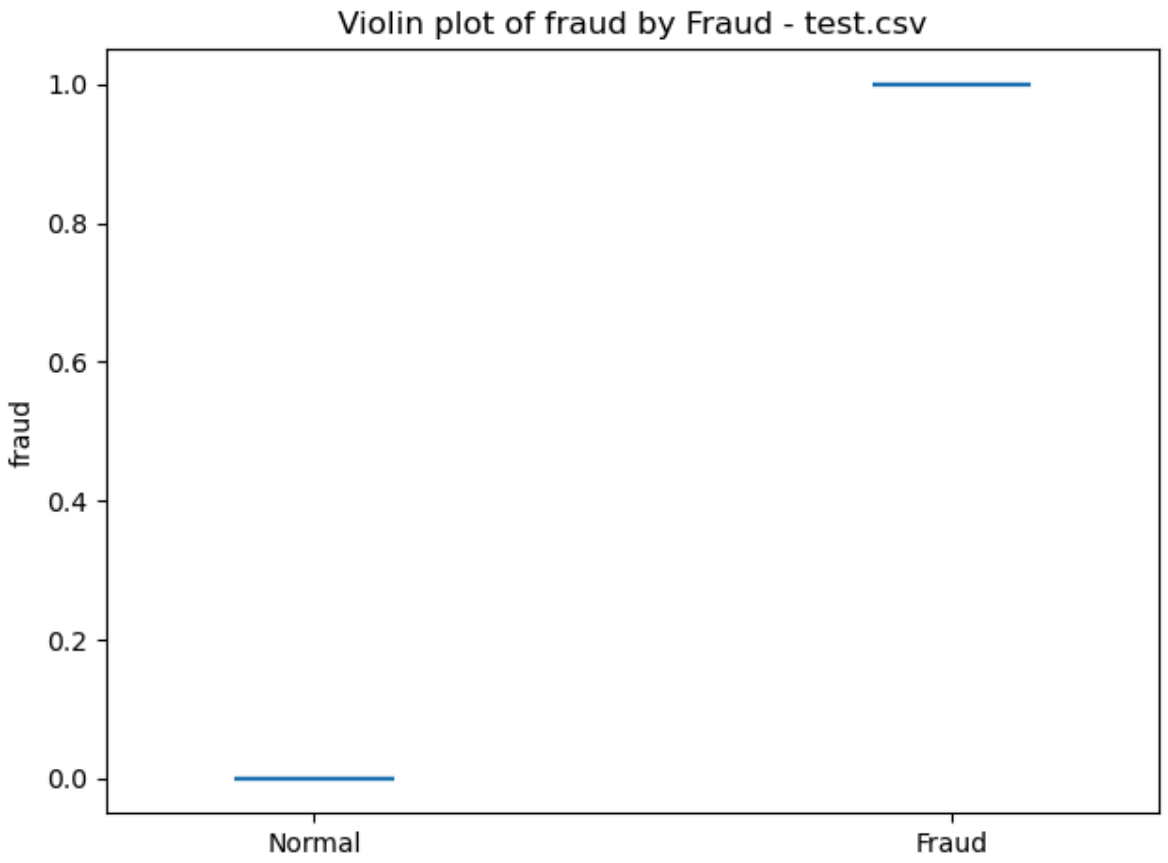
Majoritatea tranzactiilor sunt realizate pe telefon, apoi pe laptop si respectiv calulator



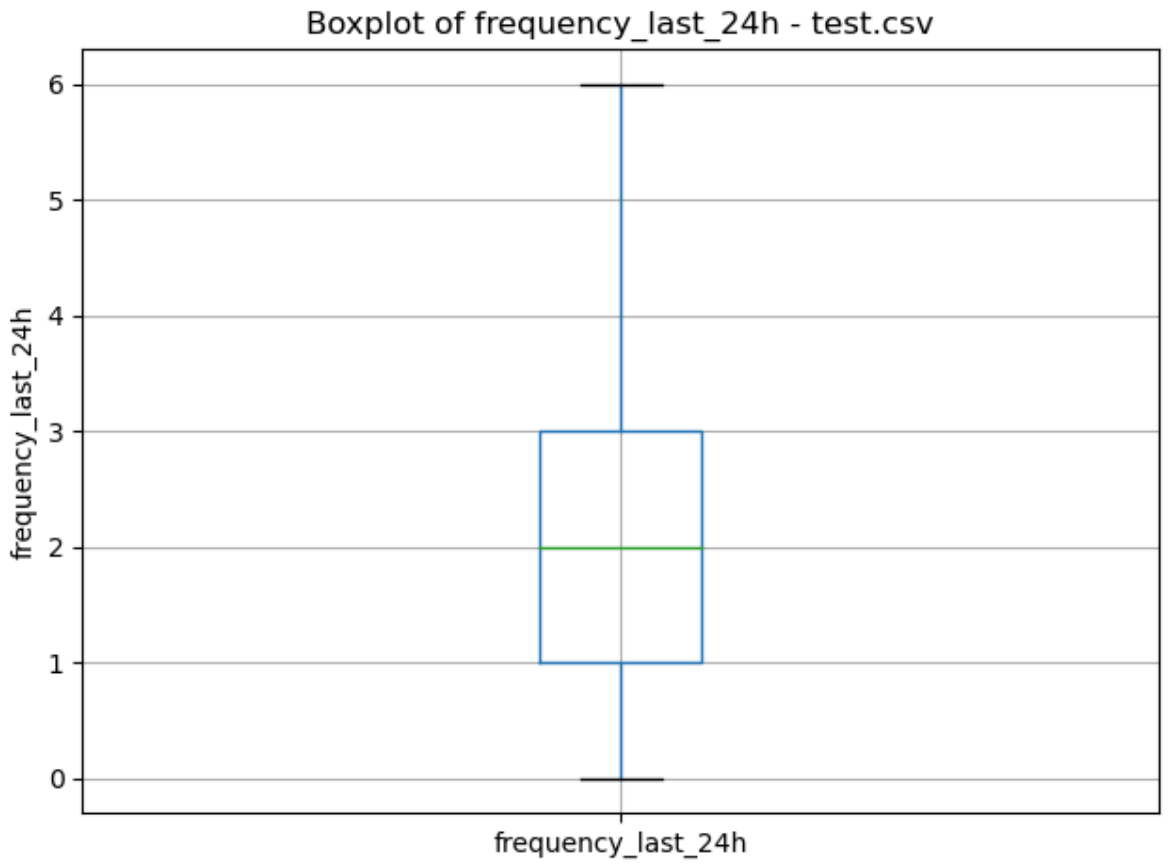
Indica ca majoritatea valorilor nu sunt frauduloase



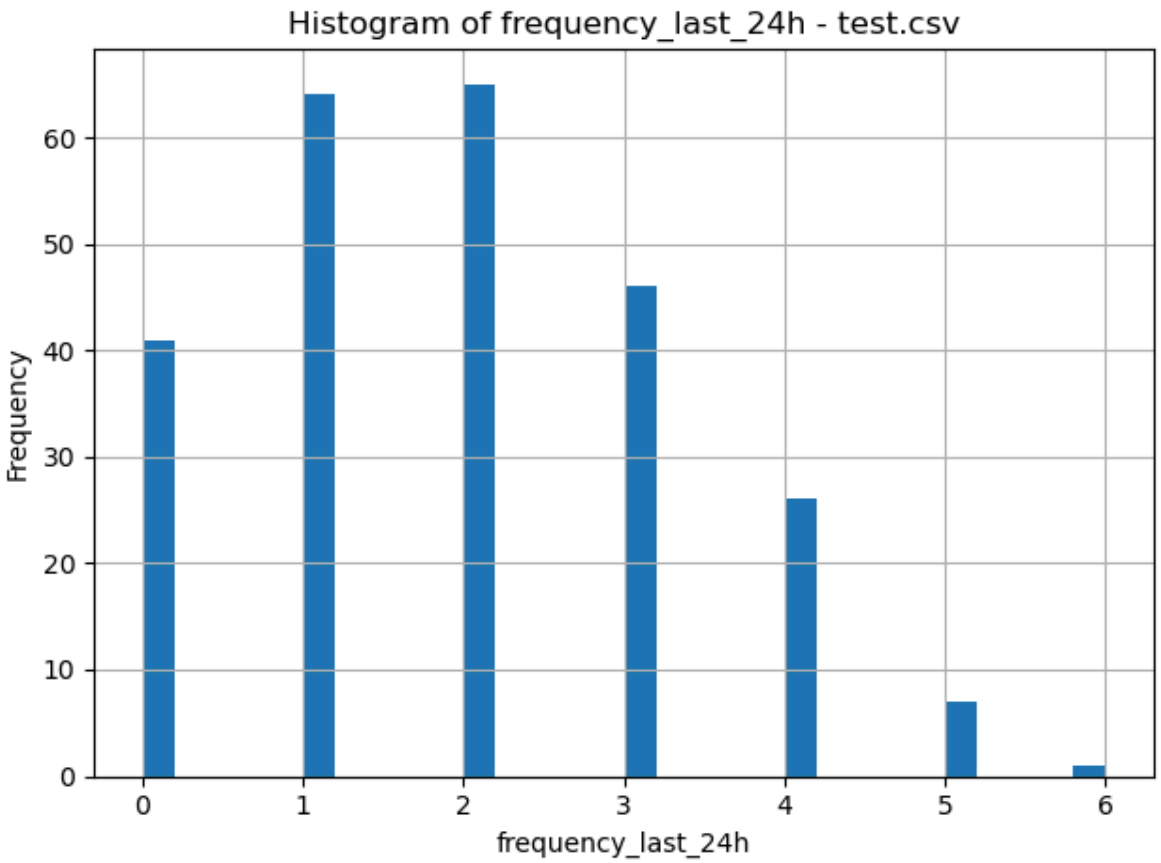
Confirma ca sunt mai multe tranzactii care nu sunt frauduloase



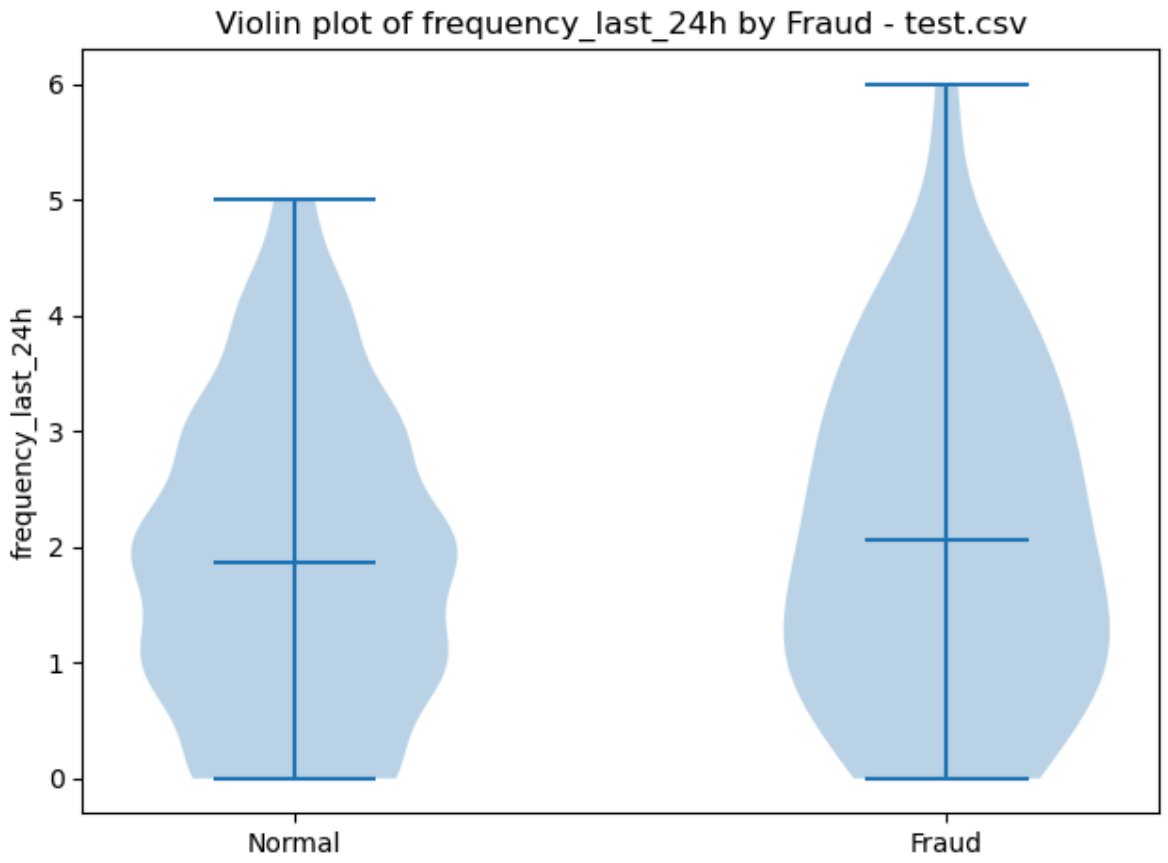
Reprezinta valorile de 0 si 1

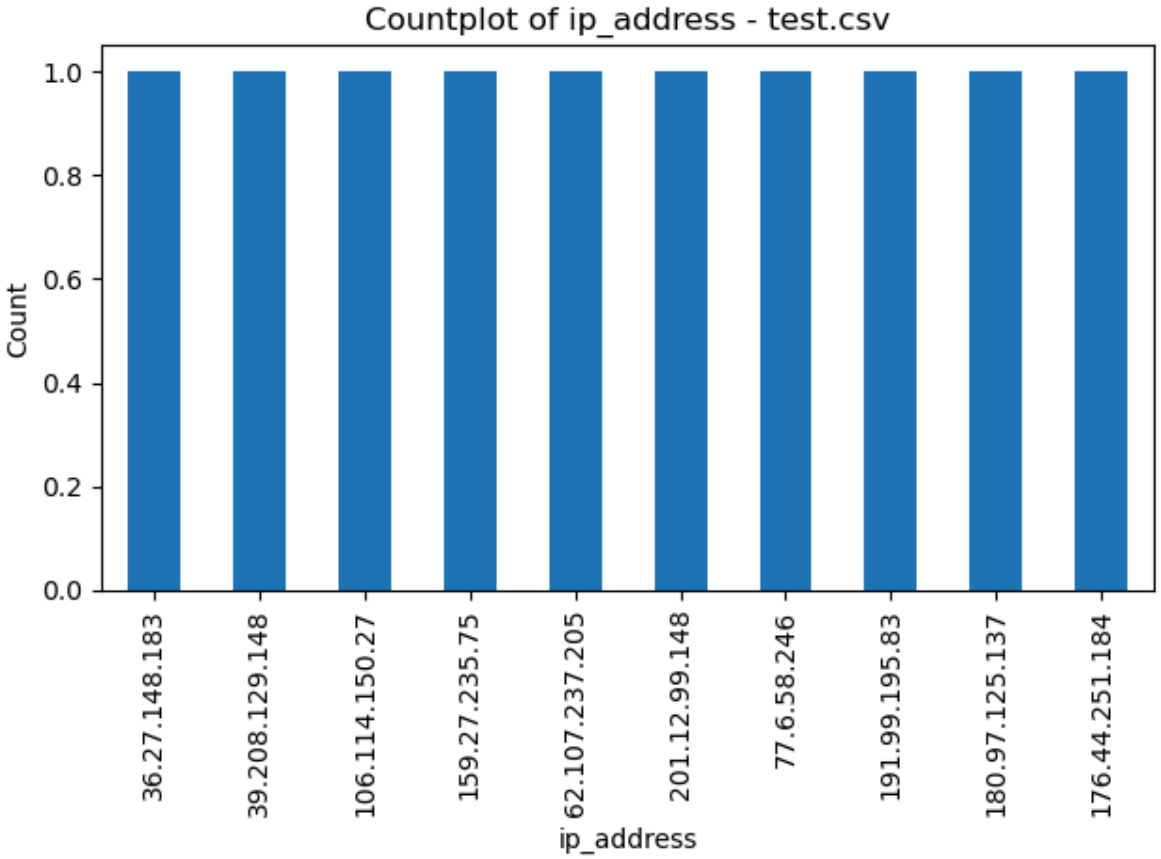


Distributia frecventei

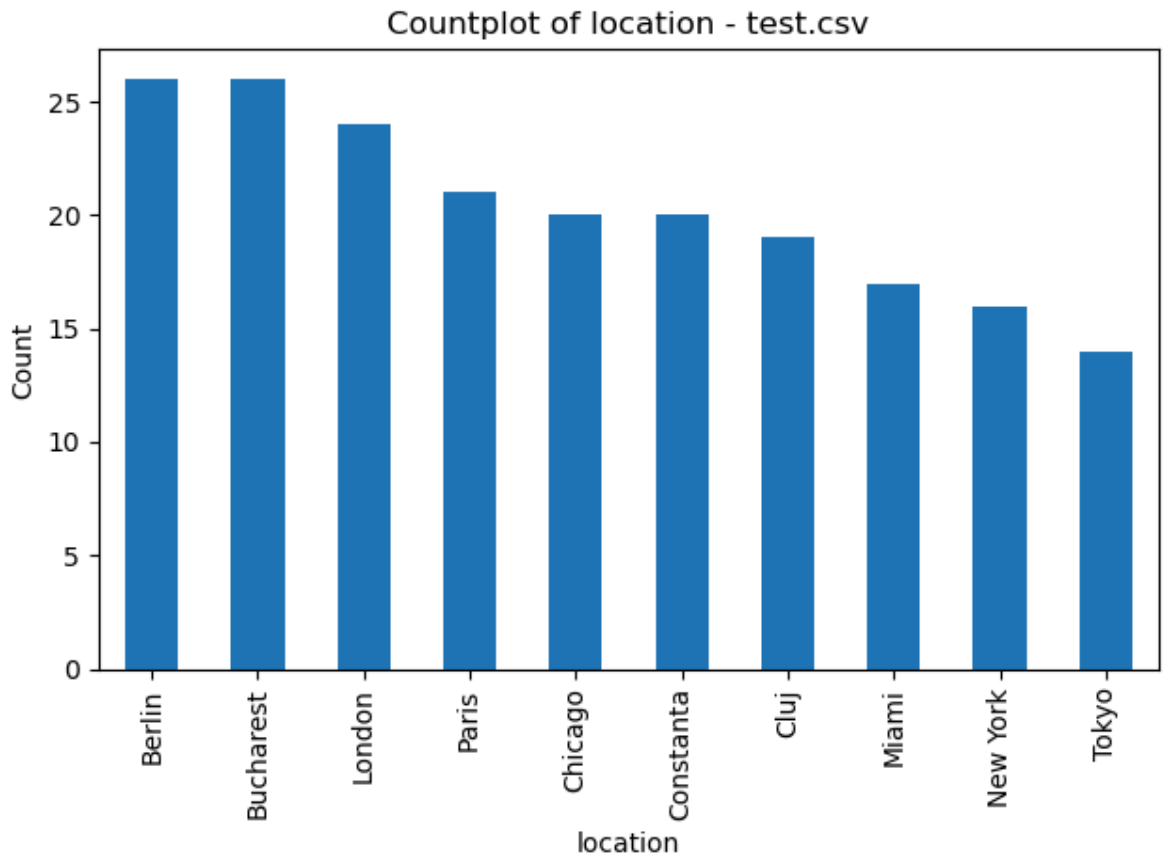


Frecventa frecventelor din ultima zi

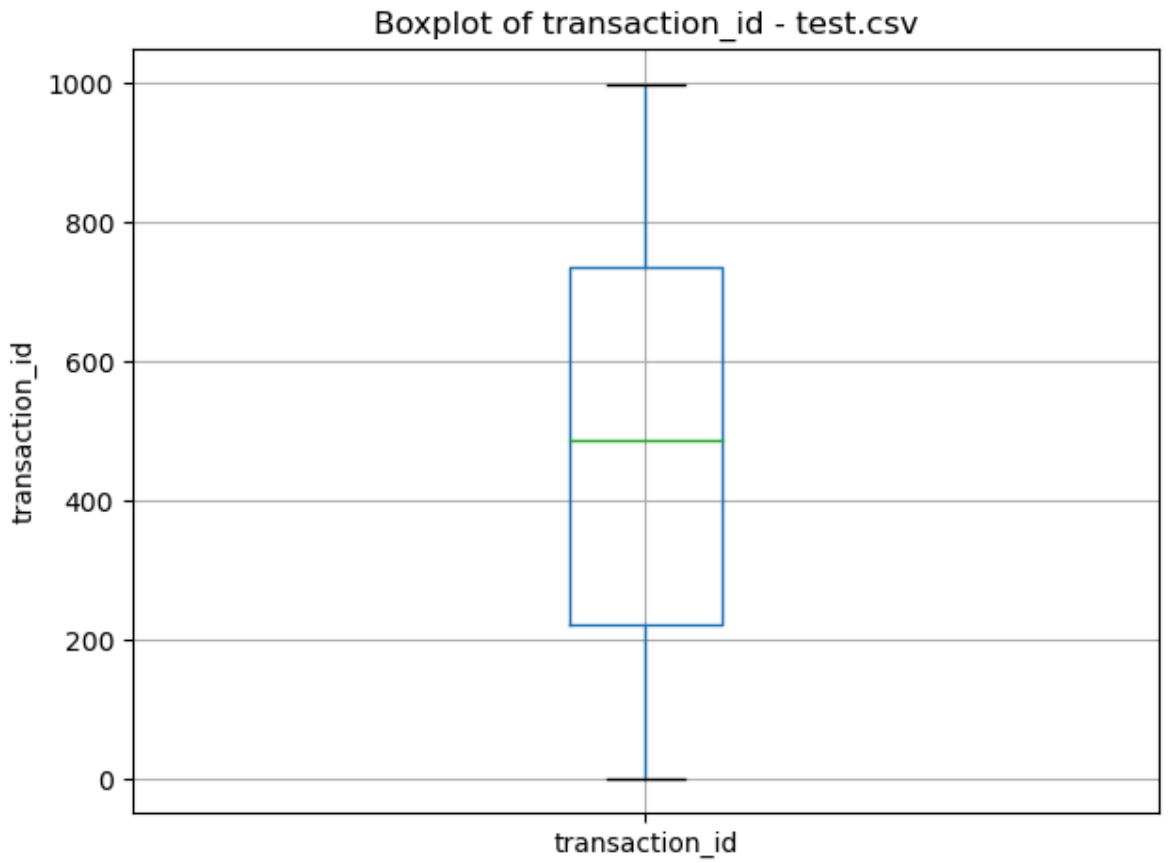

Violin plot of frequency_last_24h by Fraud - test.csv
frequency_last_24h
0
1
2
3
4
5
6
Normal
Fraud



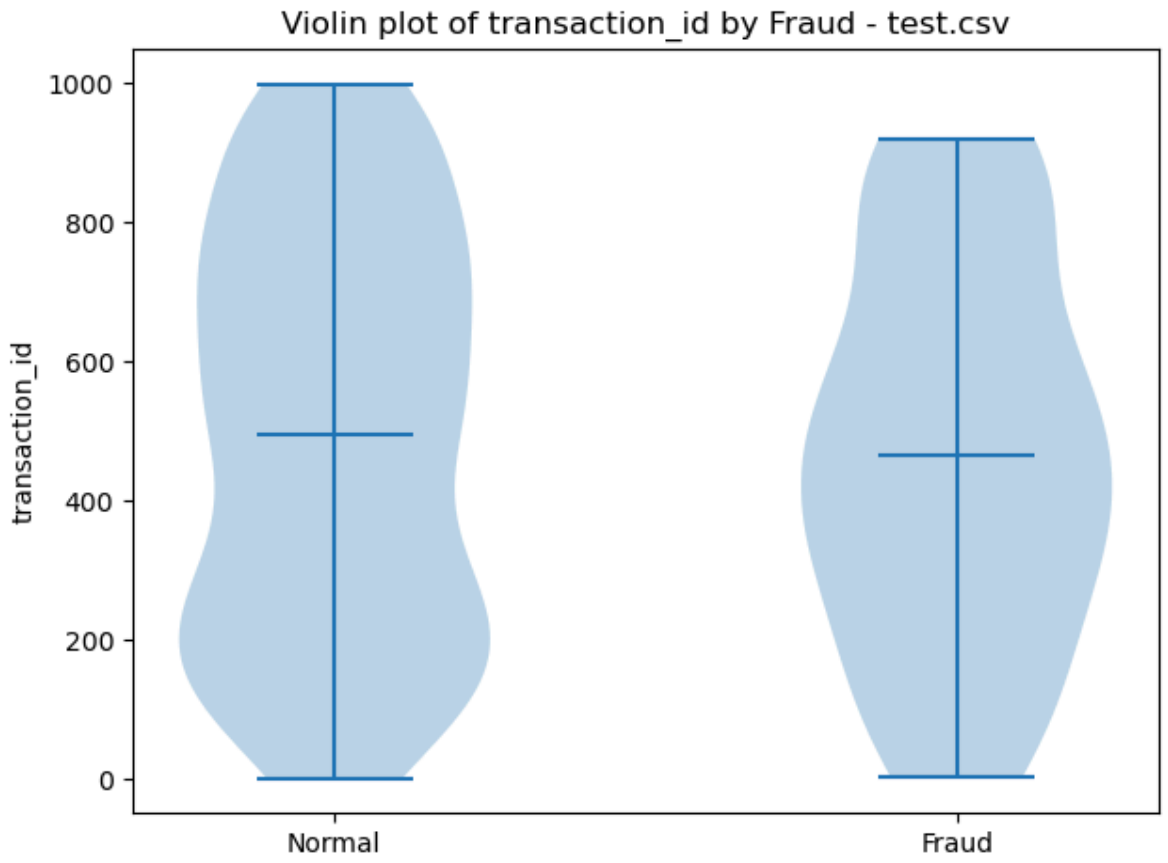
Cele mai frecvente 10 adrese ip



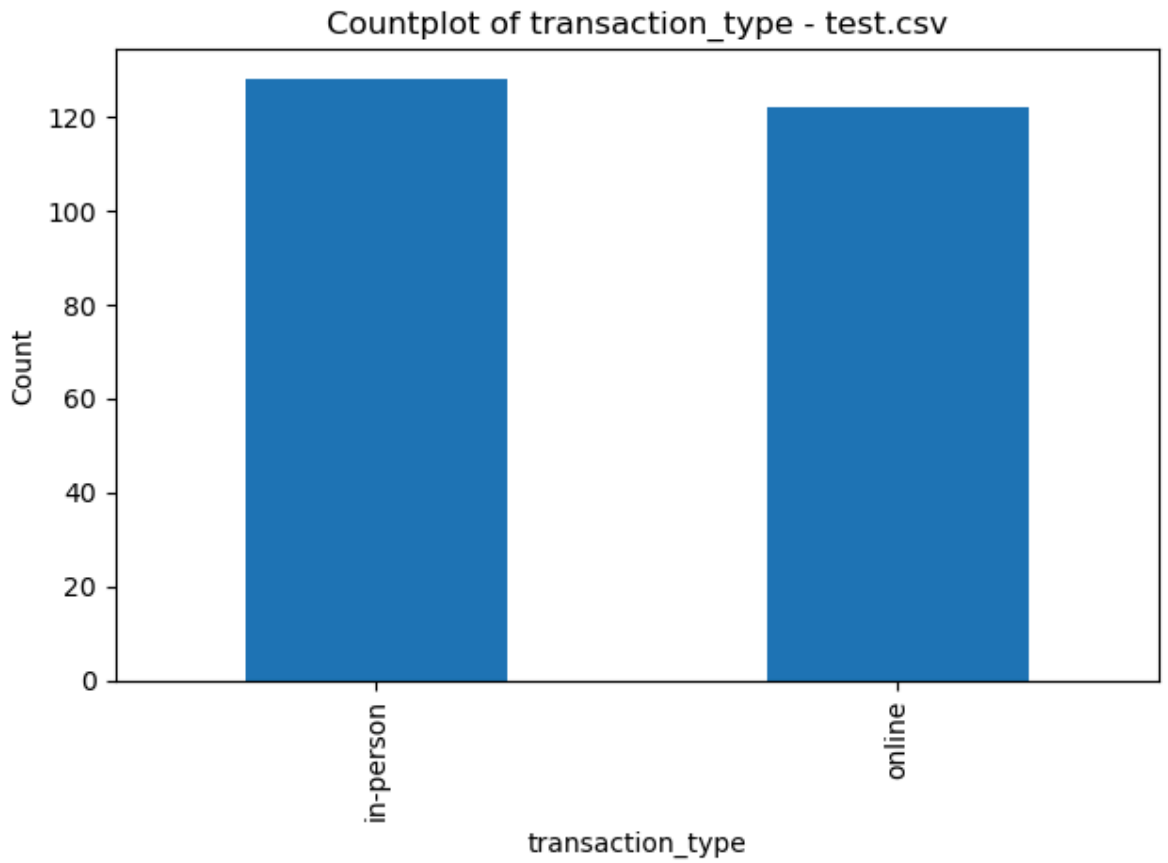
Berlin si Bucuresti au cele mai multe tranzactii



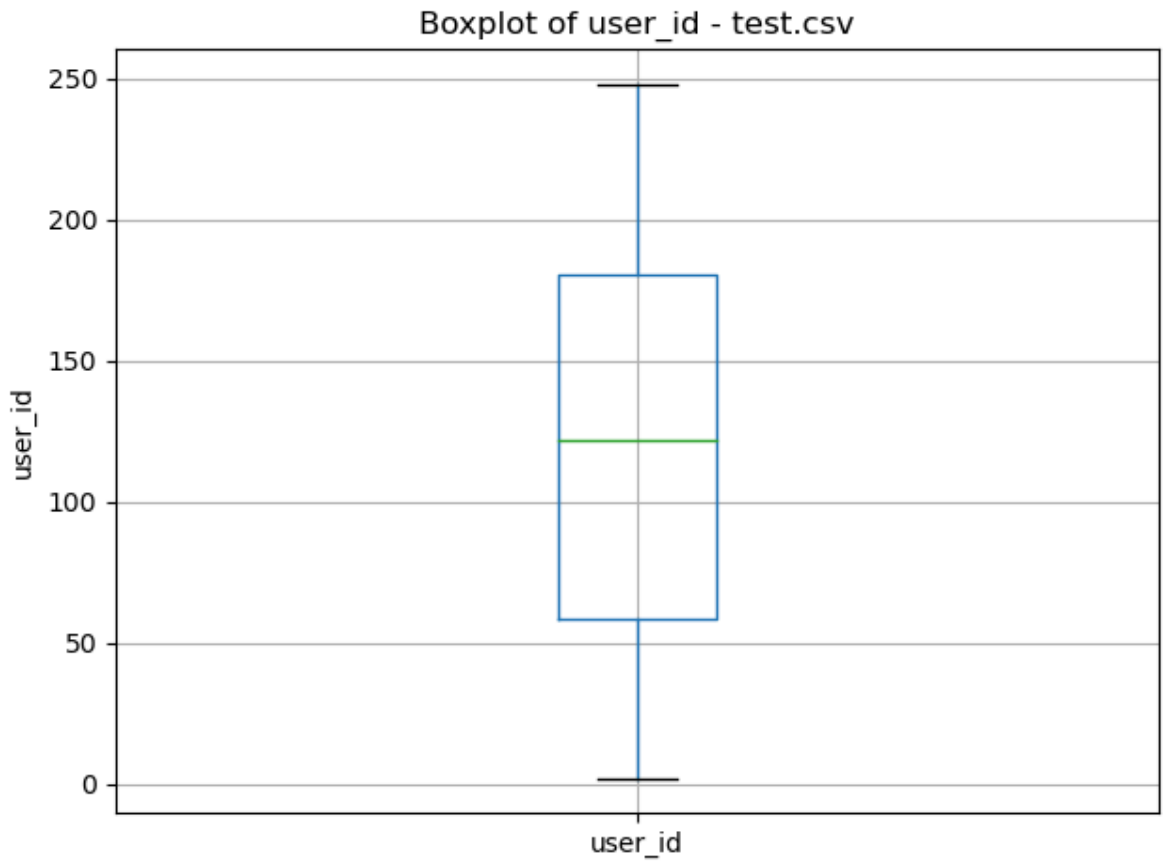
Distributia id-ului de tranzactie



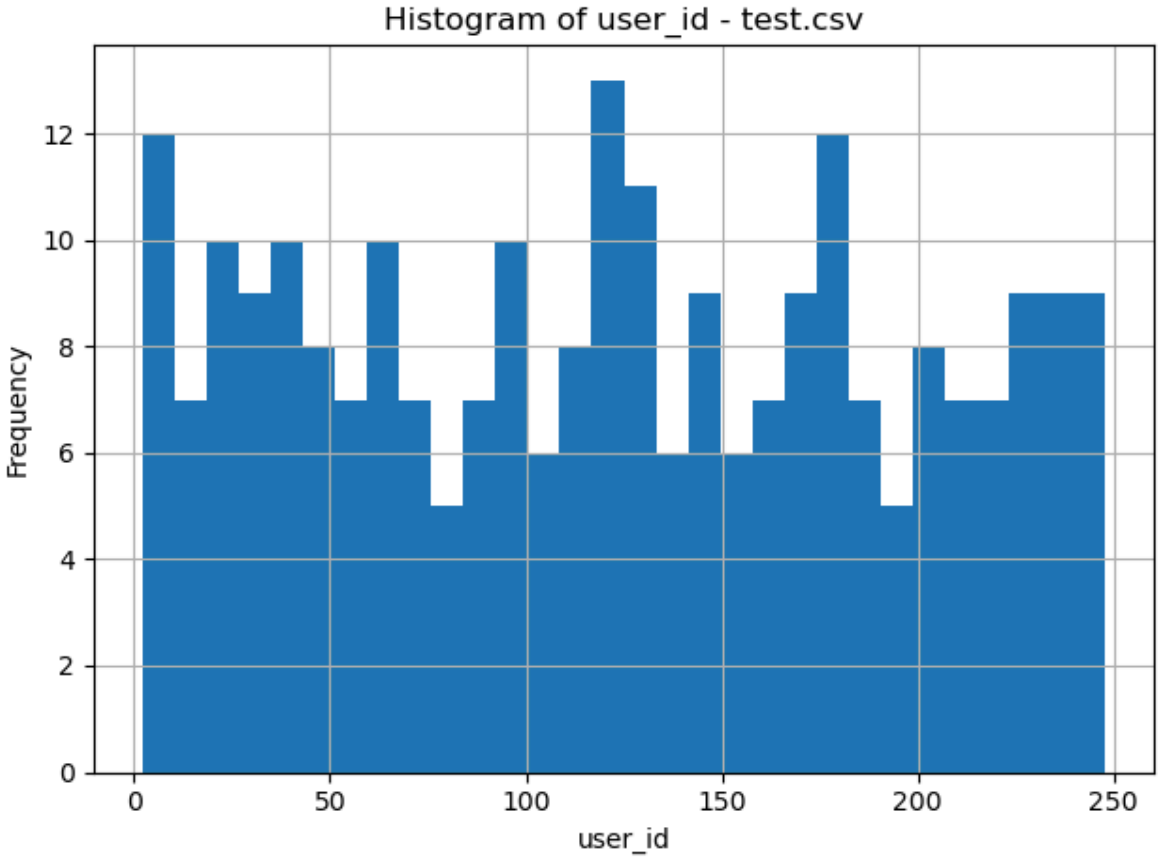
Arata ca id-ul nu este neaparat relevant



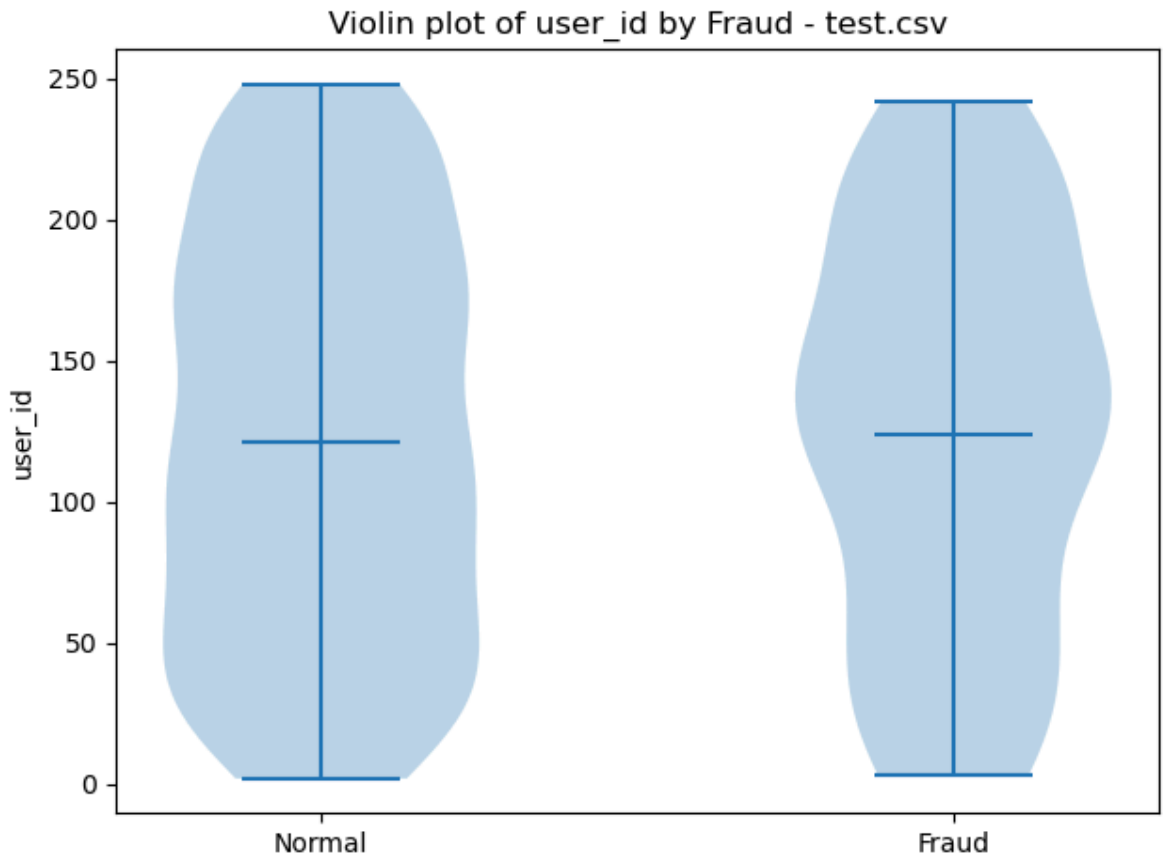
Mai multe tranzactii sunt facute in persona fizica



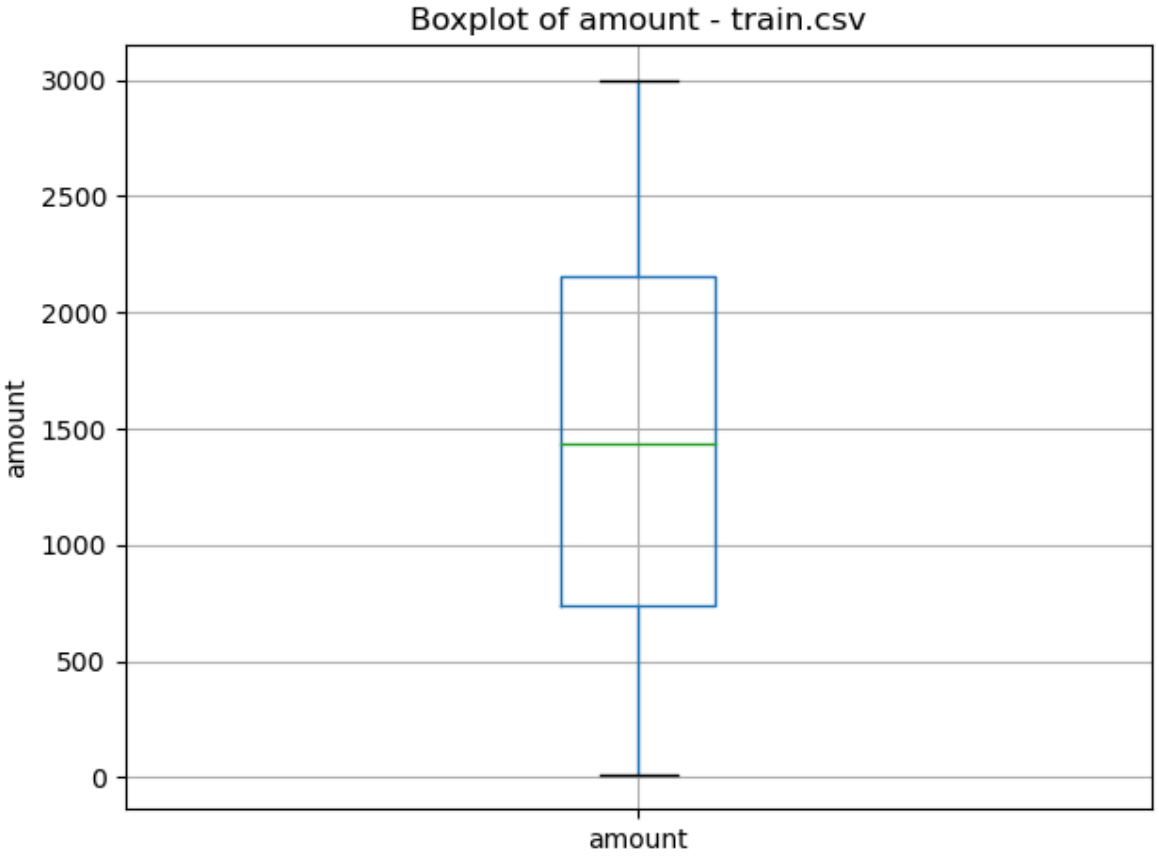
User id-urile sunt distribuite egal



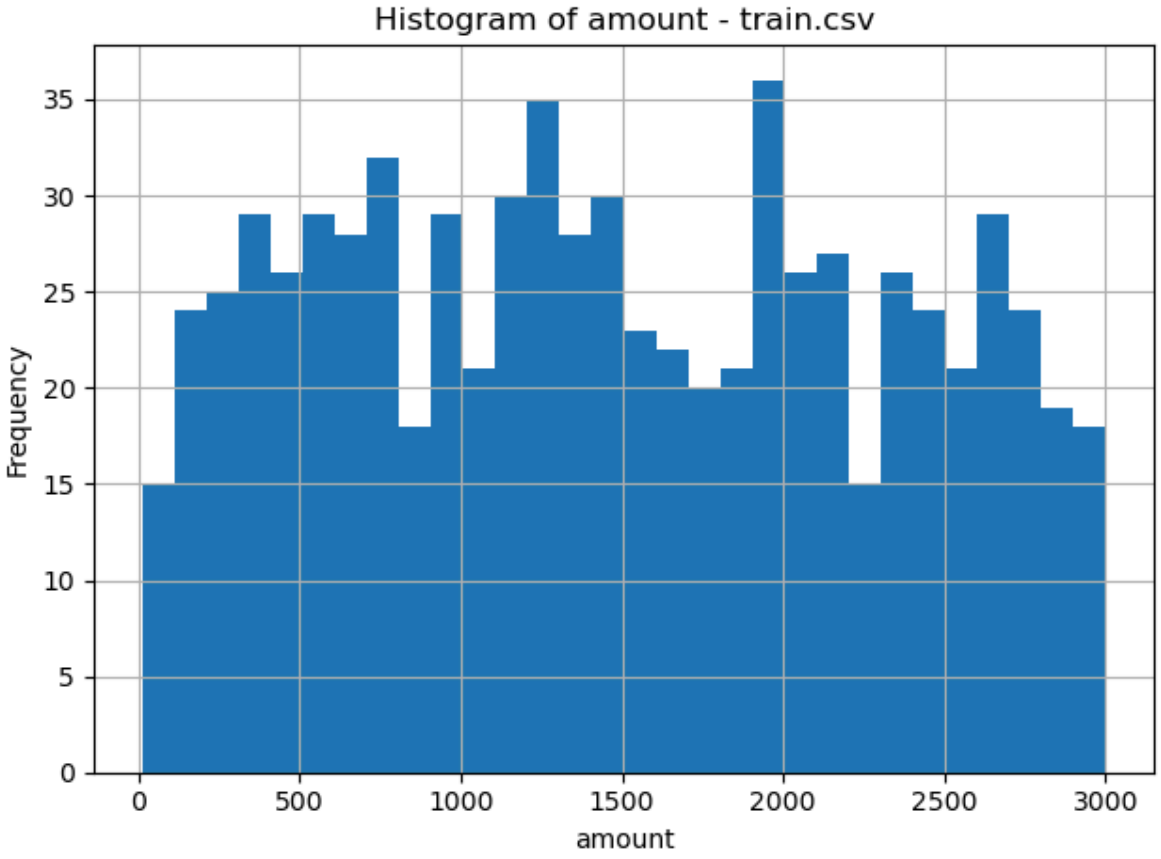
Frecventa id-urilor



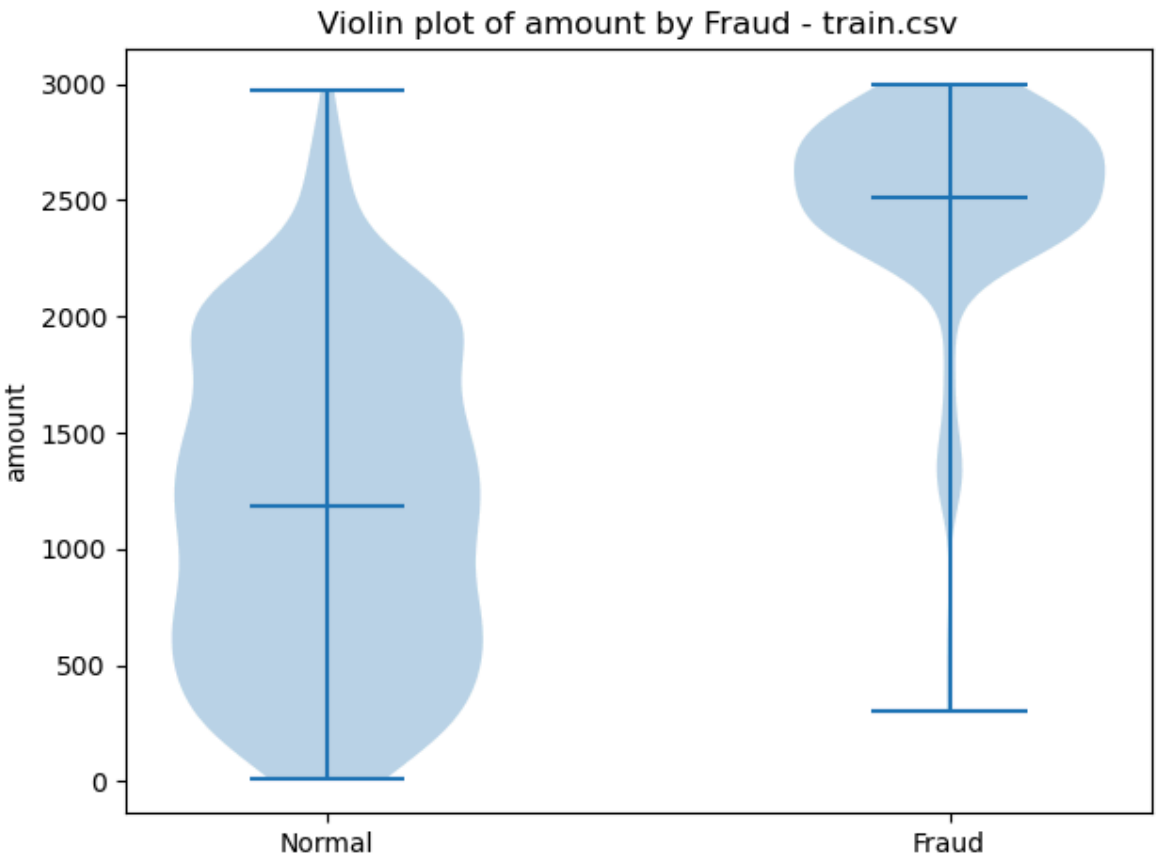
User id-urile nu reprezinta neaparat un factor important



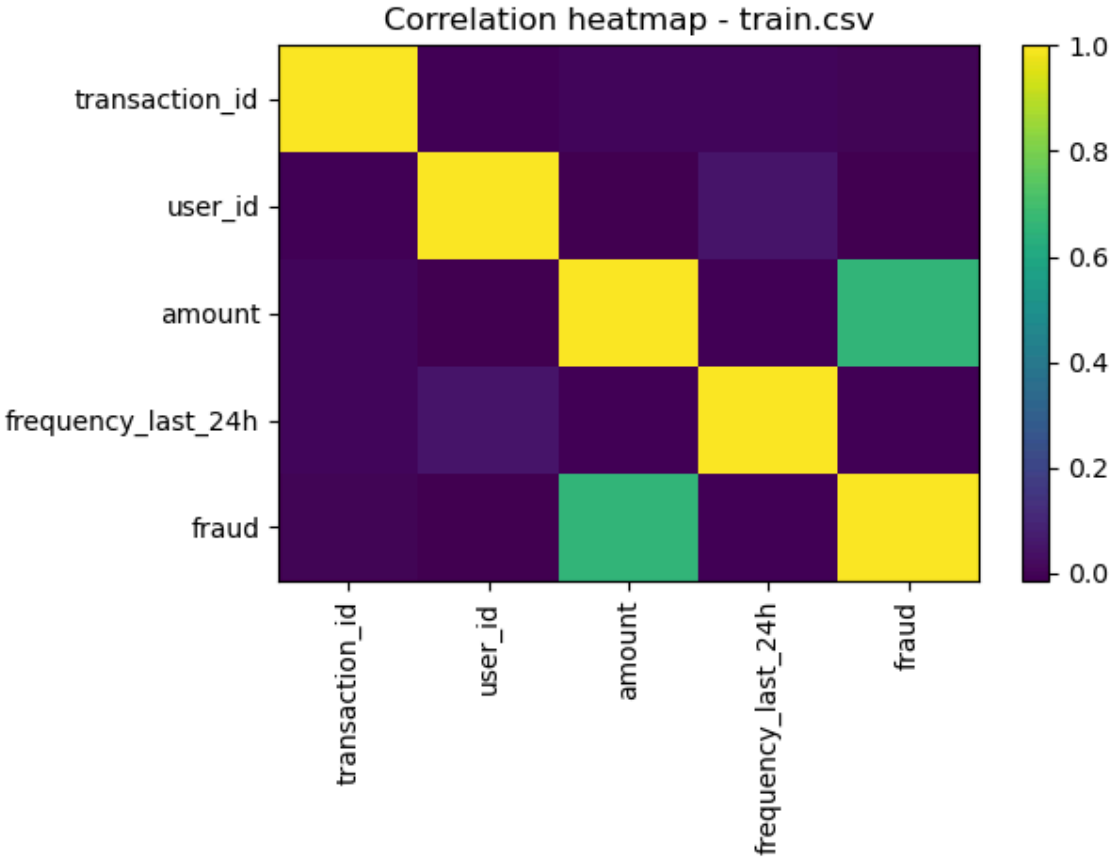
Sumele sunt distribuite destul de egal



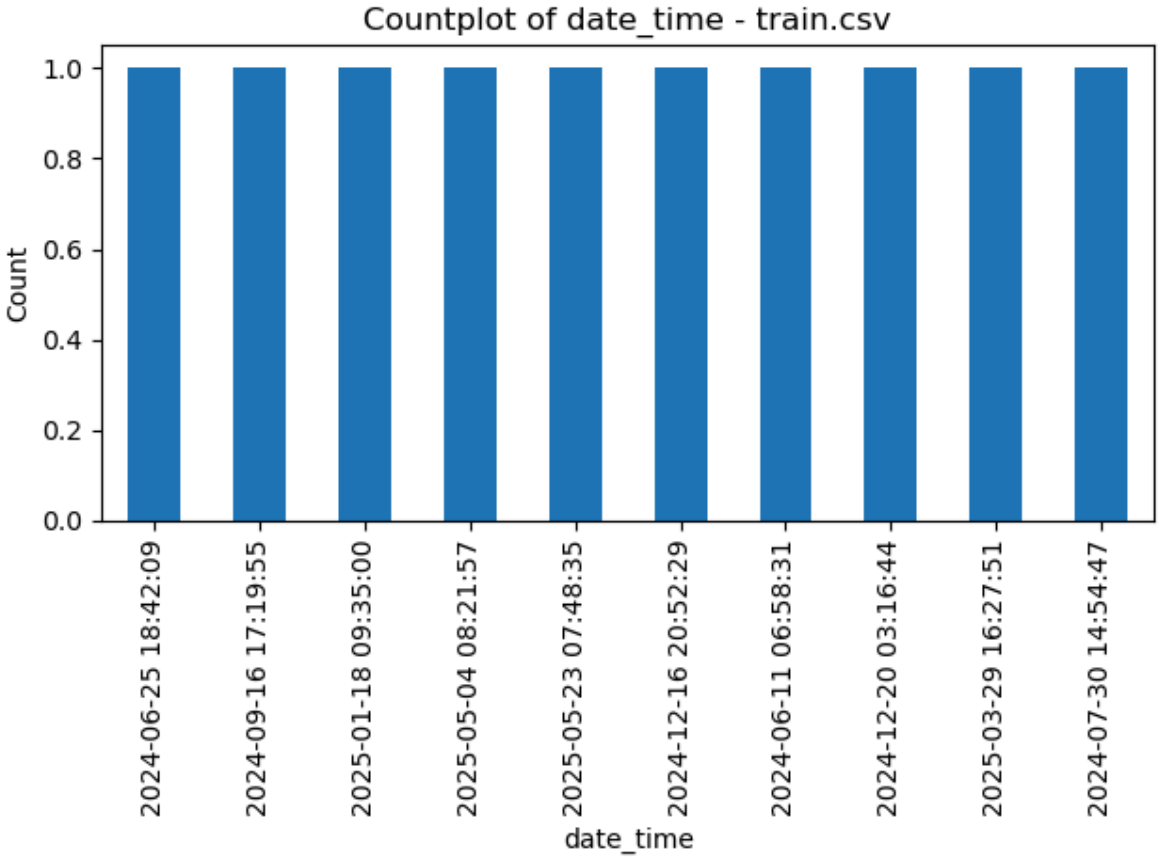
Frecventa sumelor tranzactionate



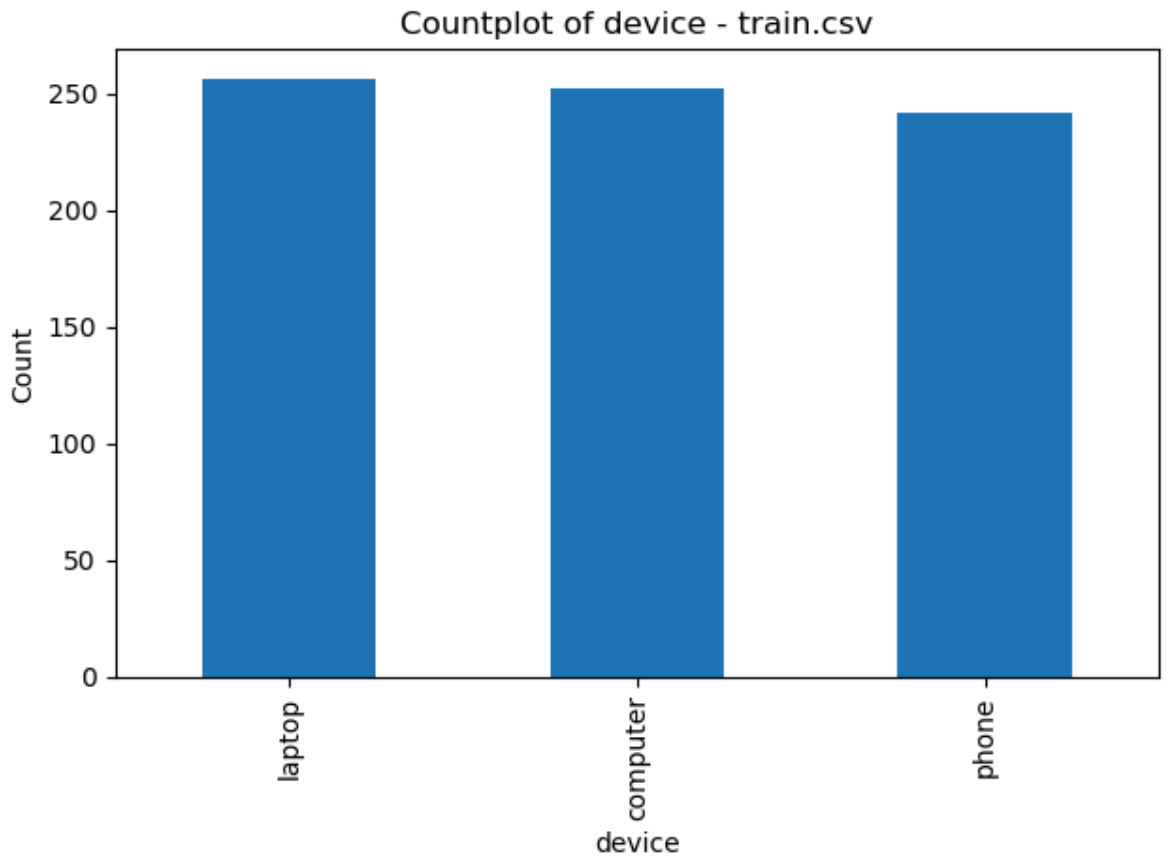
Tranzactiile frauduloase tind sa aiba sume mai mari



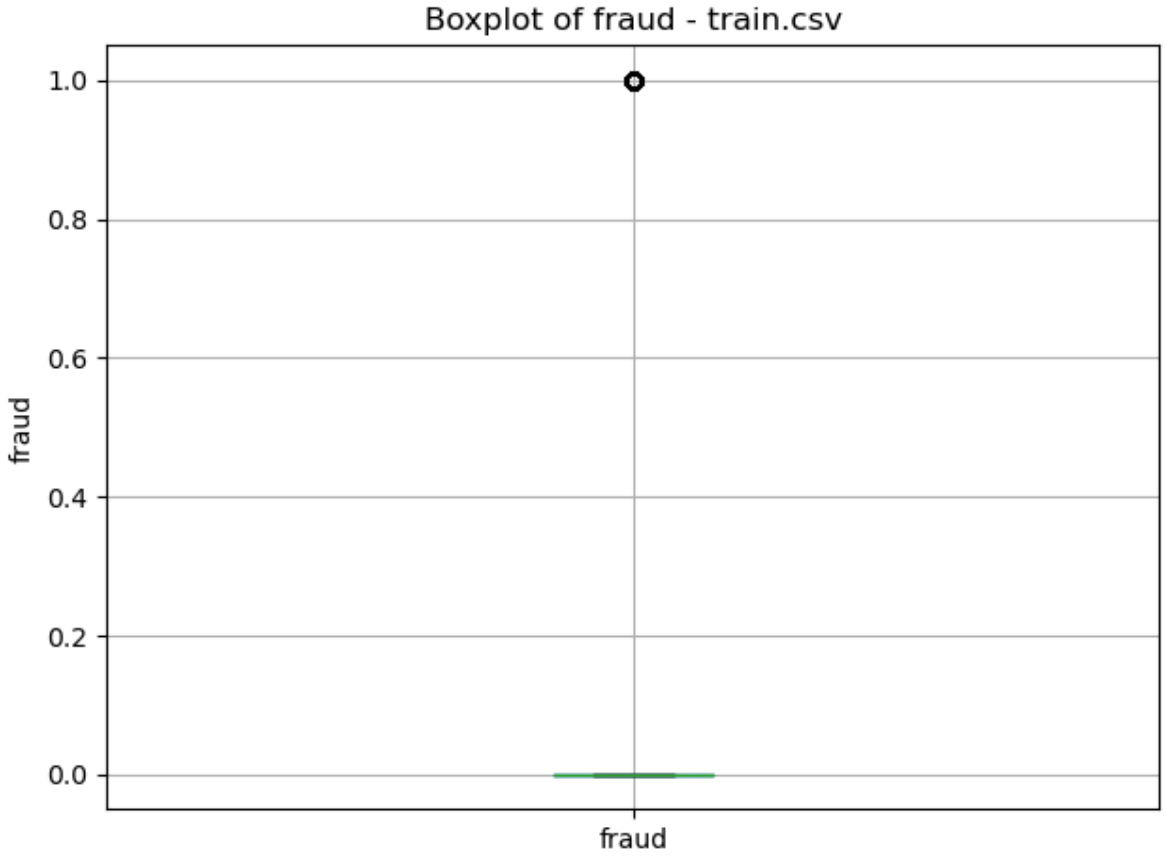
Vizualizeaza corelatiile dintre variabilele numerice in setul de antrenament



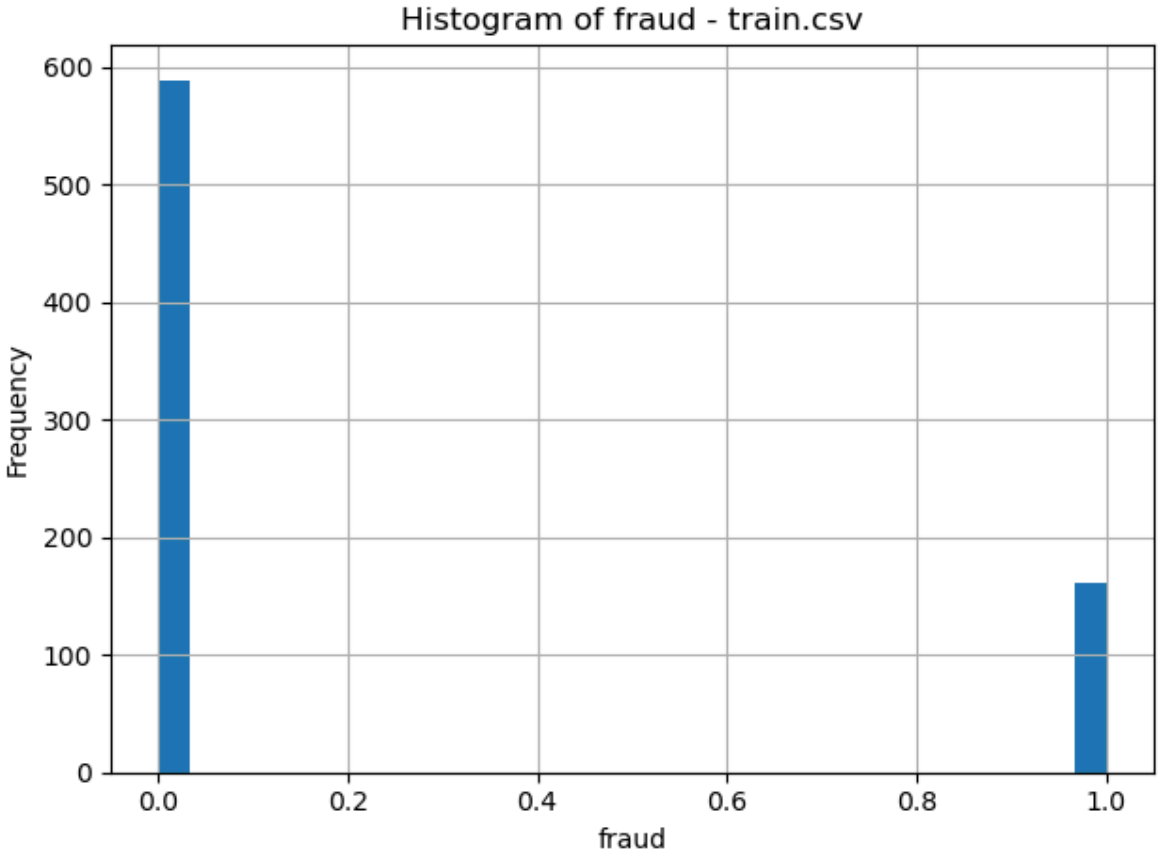
Fiecare data si ora apare o singura data



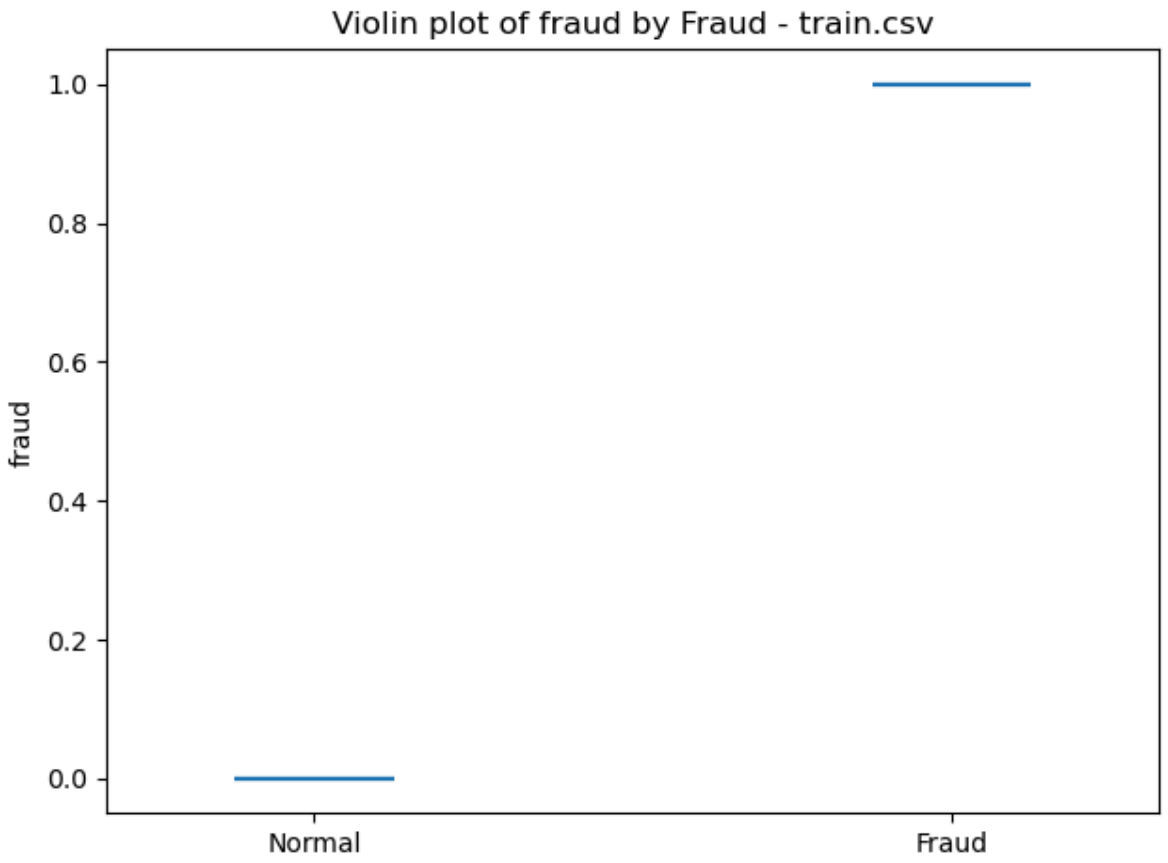
Tranzactiile sunt facute putin mai des de pe laptop decat computer sau respectiv telefon



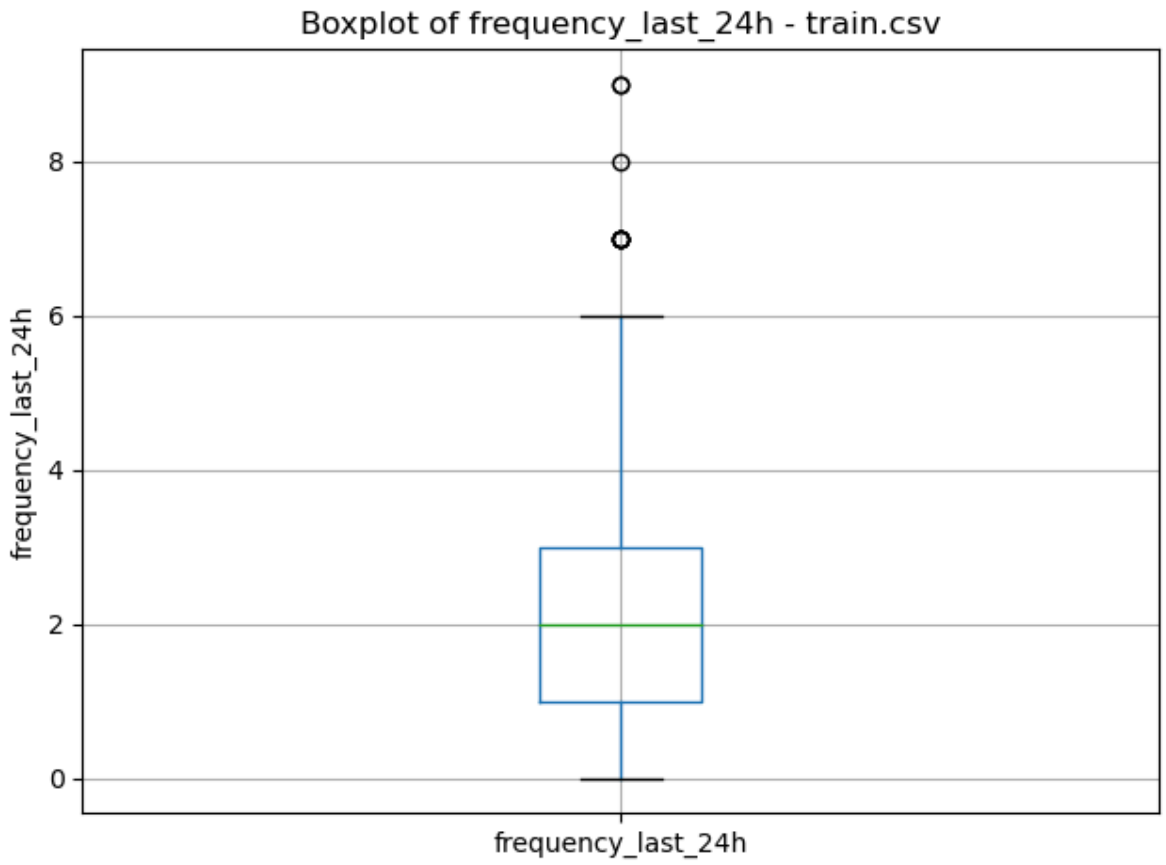
Majoritatea tranzactiilor nu sunt frauduloase



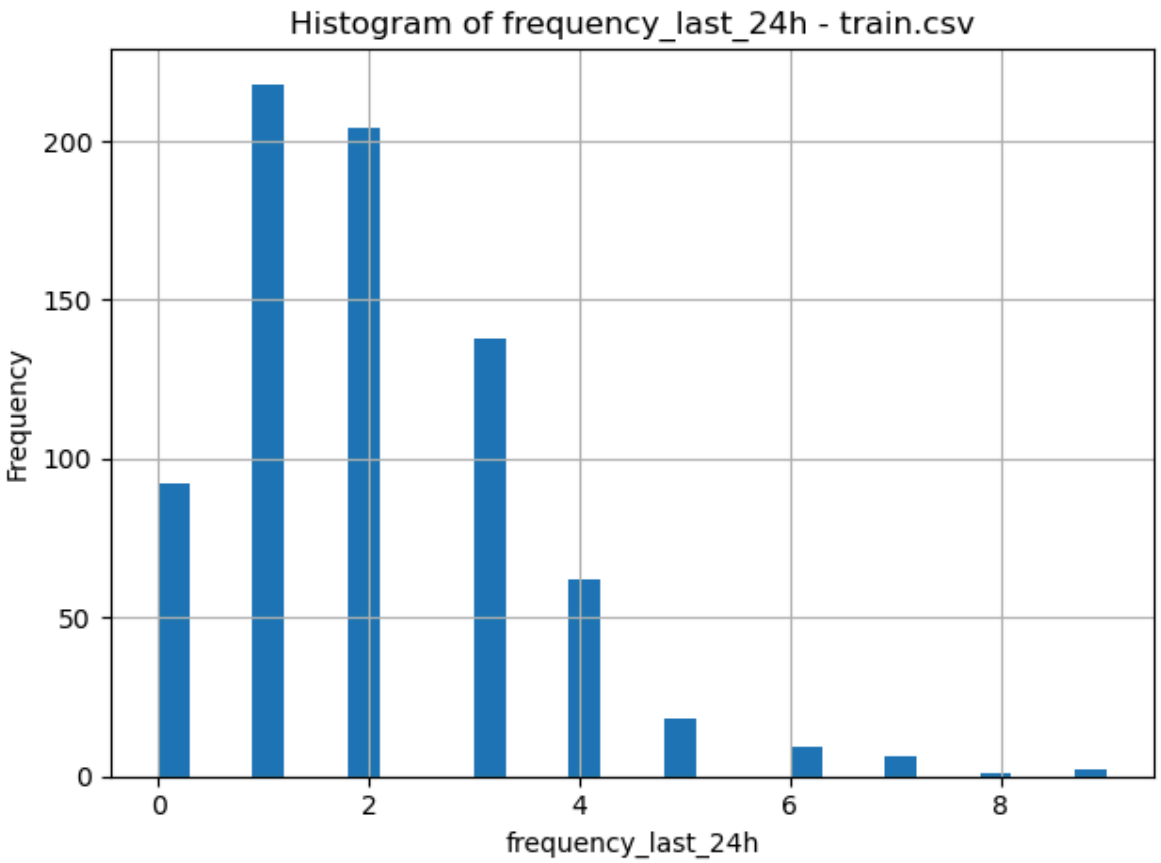
Arata ca si in setul de antrenament sunt mai multe tranzactii normale decat frauduloase



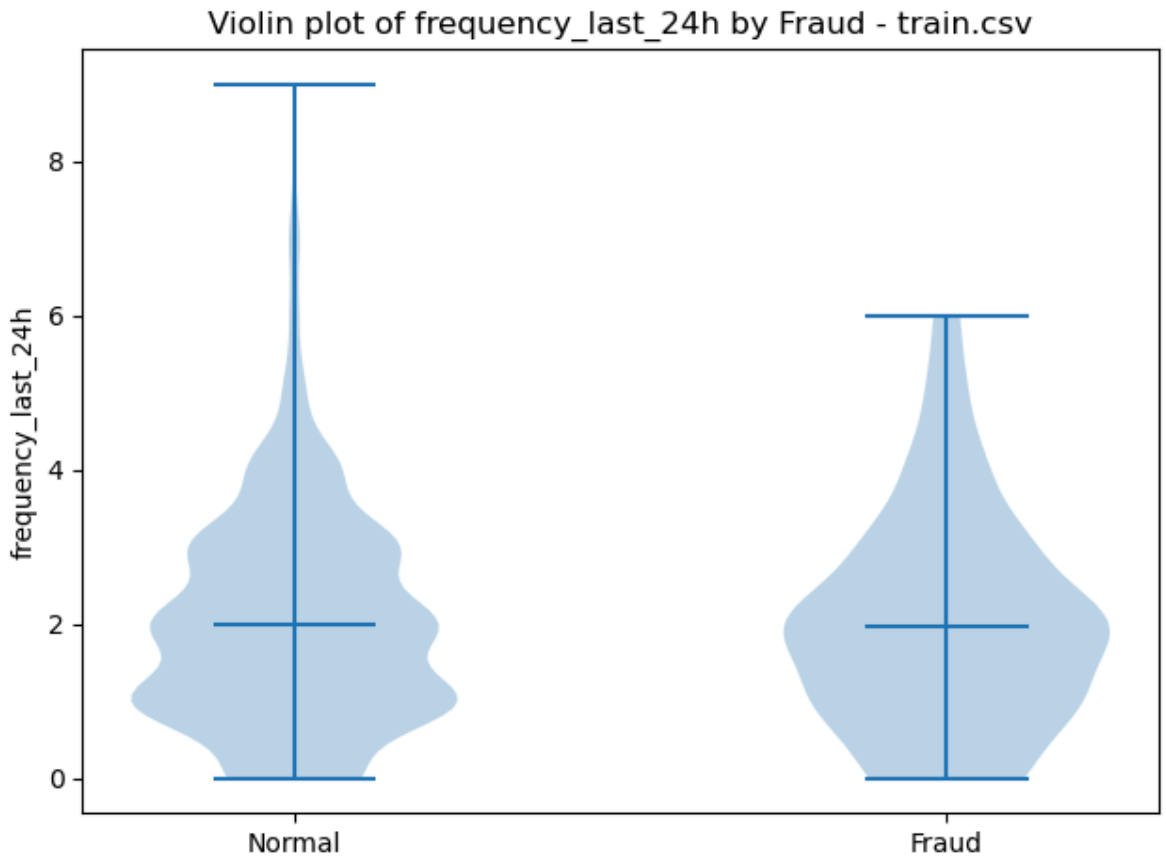
Arata 1/0 pentru frauduloase si restul



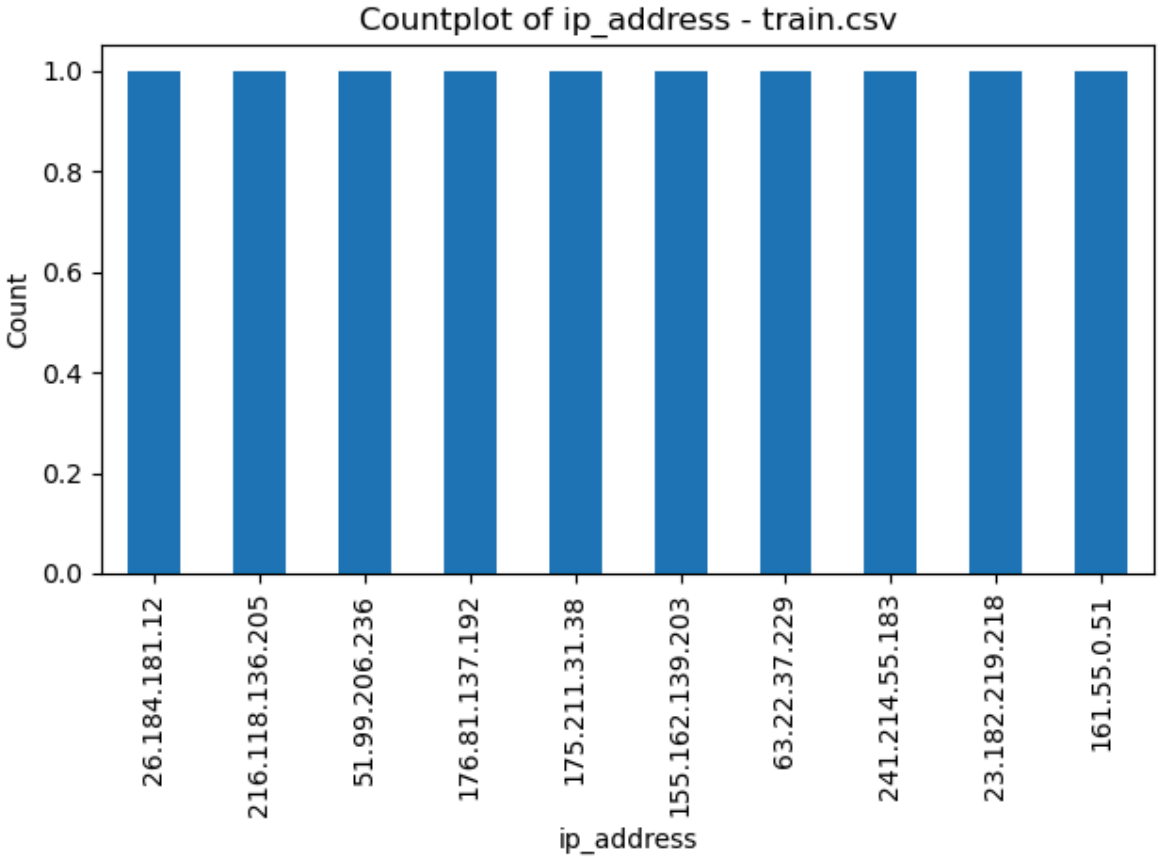
Arata distributia frecventei



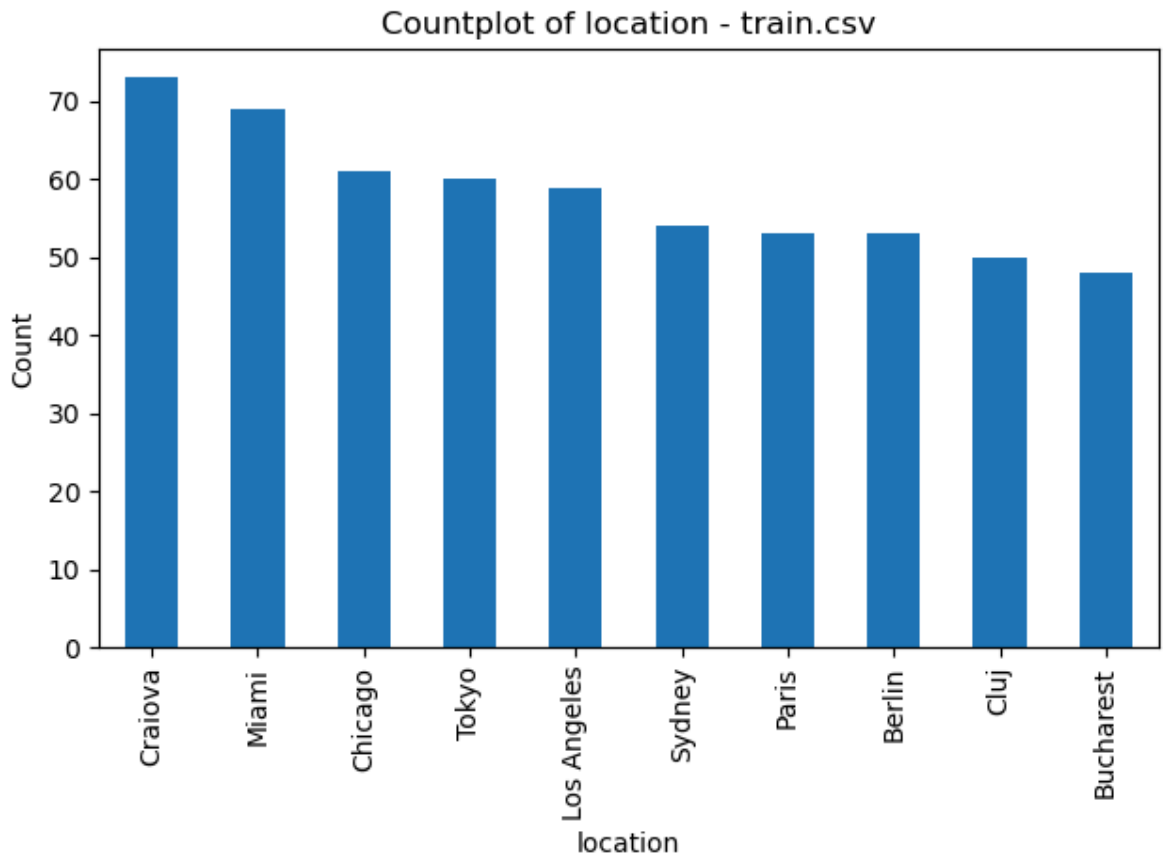
Arata frecventa din ultima zi



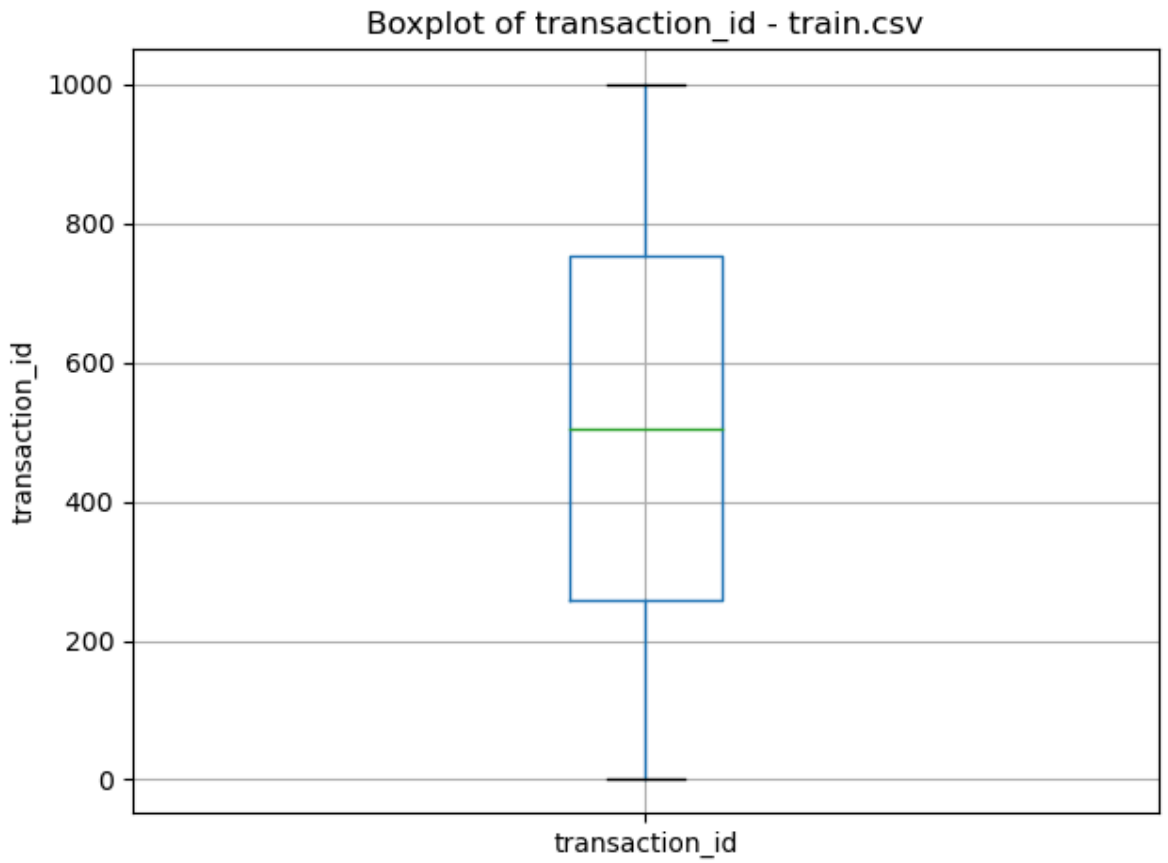
Arata ca fraudele nu sunt atat de multe zilnic



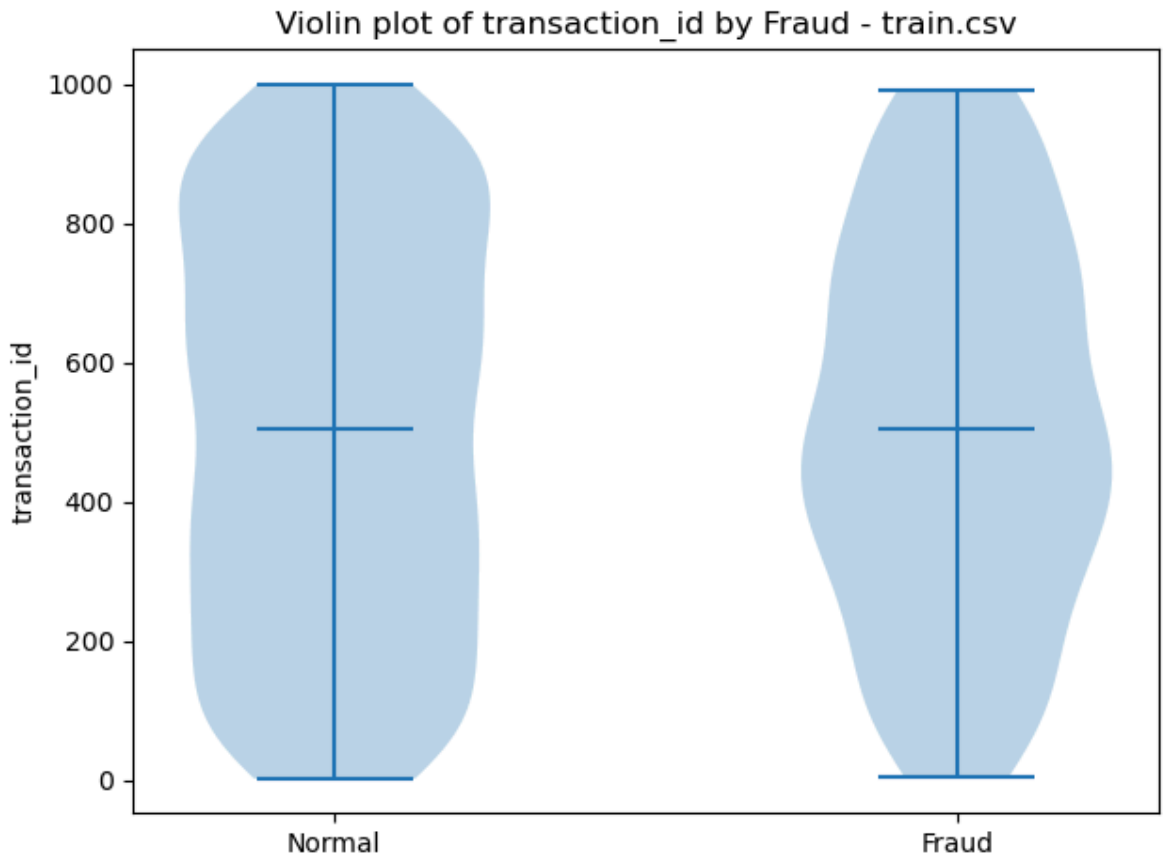
Fiecare ip are o singura tranzactie



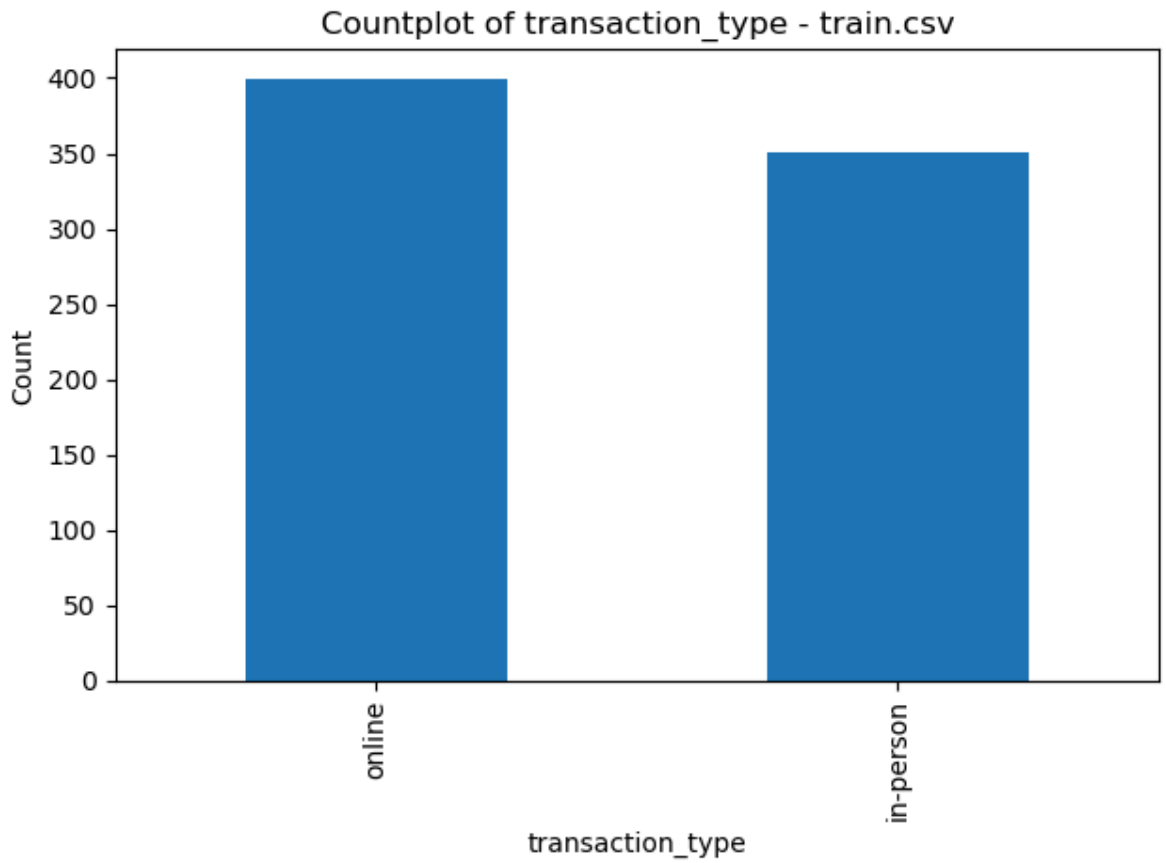
Craiova are cele mai multe tranzactii



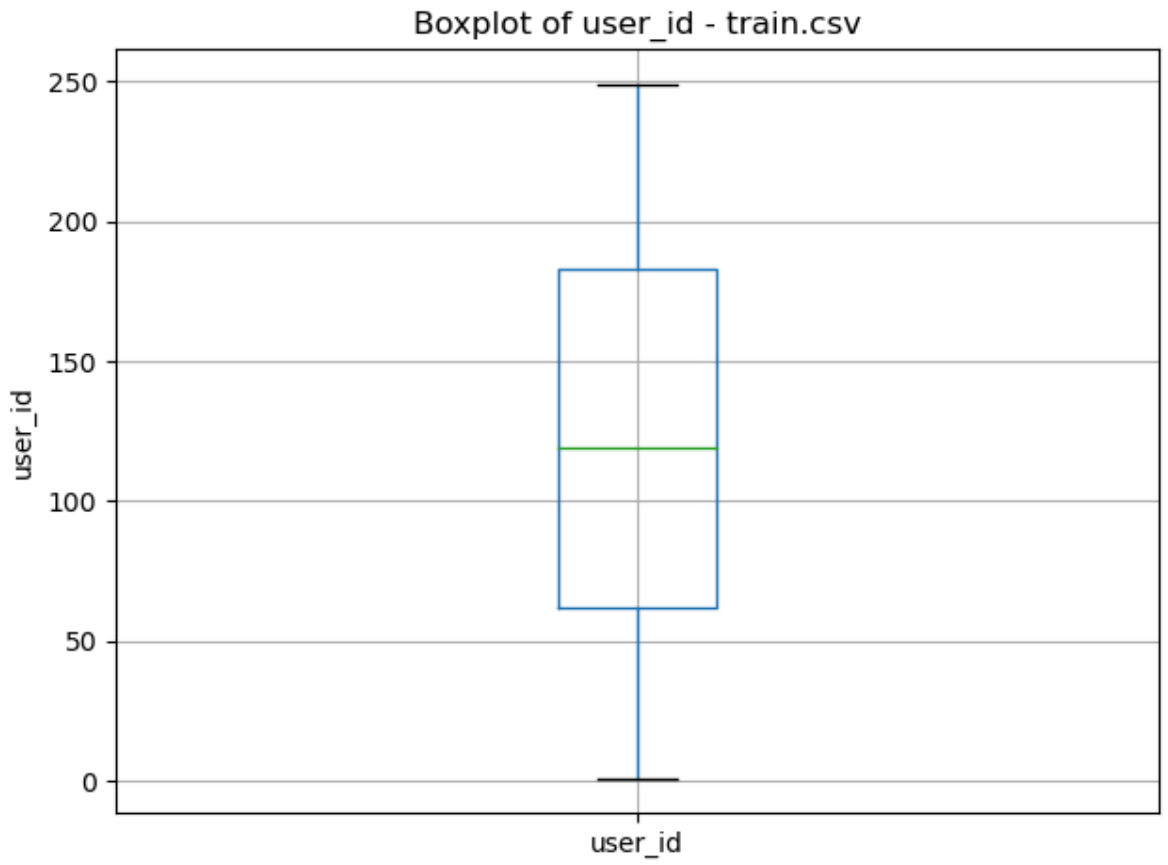
Distributia de transaction_id-uri este egala



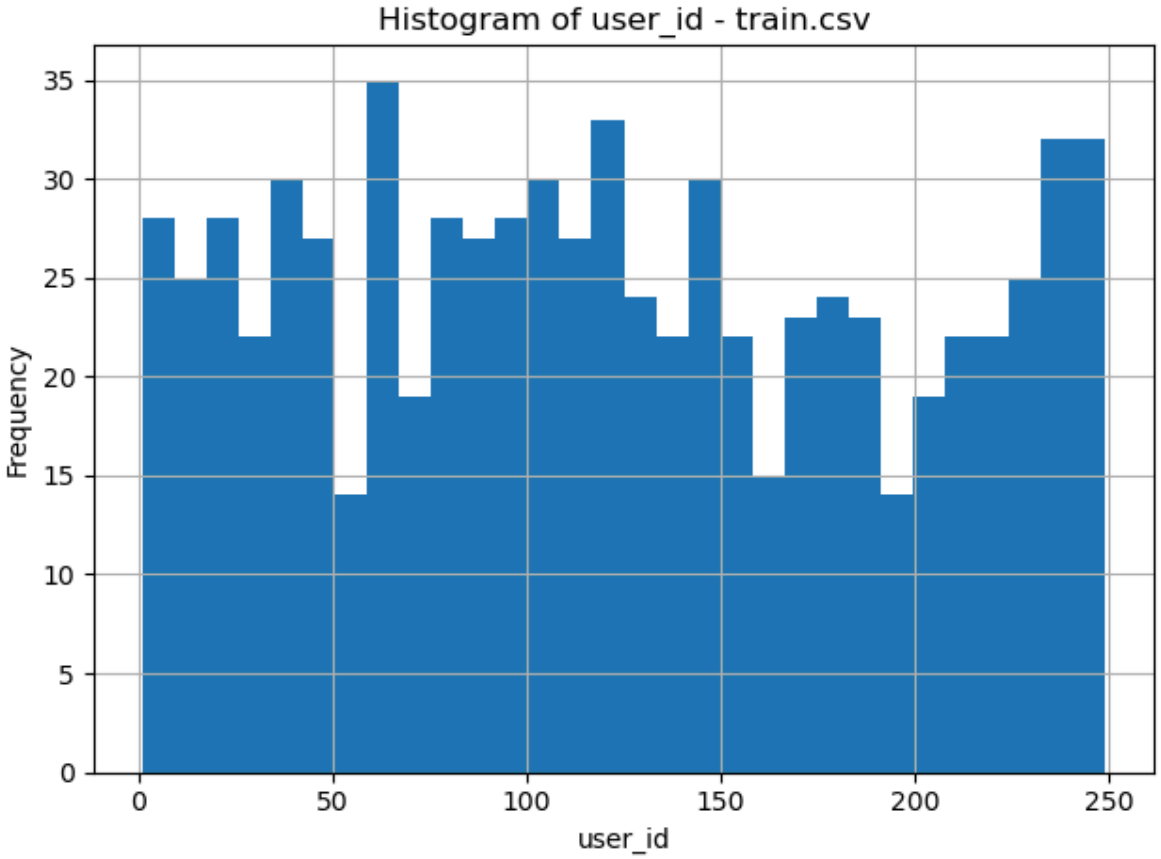
Distributia este normala



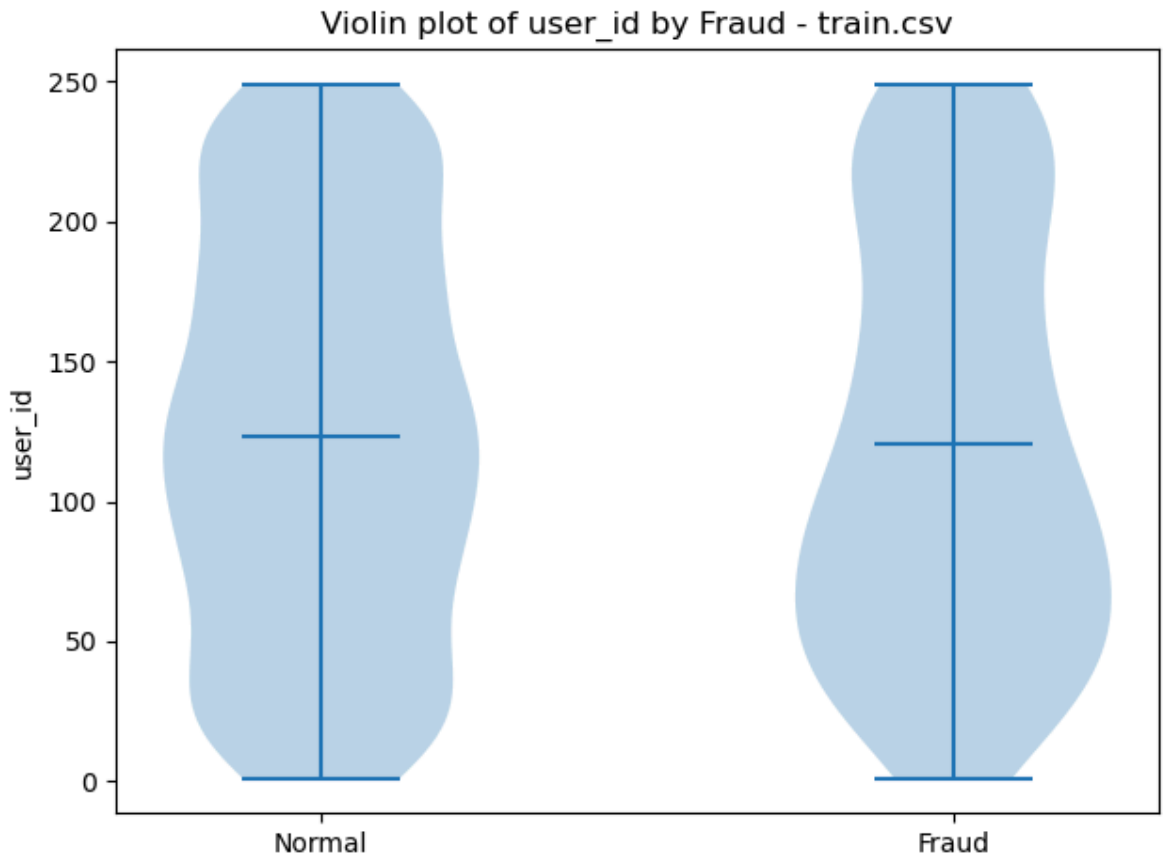
Sunt mai multe tranzactii online



Distributia de id-uri este egala



Frecventa fiecarui utilizator



Distributia este normala