Nume: Carp Andrei-Costin
Grupa: 311CC
An: I
Coleg: Solomon Stefan
Github : https://github.com/krpandrei05/Tema1_PCLP3.git

Partea I - Construirea si explorarea unui dataset tabelar

Luand in considerare cele 3 metode prezentate pentru a construi un set de date tabelar si citind despre fiecare in parte, am hotarat, impreuna cu Stefan, colegul de echipa, sa alegem a doua metoda
propusa (generarea sintetica a unui dataset).

1. Tipul problemei
   -> Problema de clasificare
   -> Am ales acest tip de problema, deoarece ofera rezultate mai usor de interpretat decat problema de tip regresie (grafice mult mai clare).

2. Structura setului de date
   -> Am stabilit, pentru o antrenare mai buna sa aleg 1000 de de instante:
      *Subset de antrenare: cel putin 500 => 700 (0.7 din 1000)
      *Subset de testare: cel putin 200 => 300 (0.3 din 1000)

3. Numarul minim de caracteristici
   -> Am ales ca pentru fiecare instanta sa am 10 coloane (inclusiv coloana tinta).
   -> Am ales 3 tipuri diferite de date (int, float, valori categorice).
   -> Mai multe despre fiecare coloana in sectiunea "5. Documentare".

4. Salvarea dataseturilor
   -> Am impartit dataset-ul in 2 categorii:
      *Subset-ul de antrenare (700 instante)
      *Subset-ul de testare (300 instante)

5. Documentare
   -> Setul de date simuleaza comportamentul a 1000 de clienti online. acesta fiind descris printr-o serie de caracteristici (coloane).
   -> Datele le-am generat sintetic folosind diferite functii din "Numpy" aleatoare, cu ajutorul unui seed fix (2025) pentru reproductibilitate.
   -> Datele pe care i le-am acordat fiecarei coloane am incercat sa fie destul de reale.
   -> Caracteristici generale (9 coloane):

Variabila	**Tip**	**Descriere**
valoare_comanda_medie	Numerica	Generata cu distributia normal (Gaussiana), media 400, deviatia 200, limitata intre 10 si 1000, cu 2 zecimale fiecare nr.
aplicatia	Categorica	Foloseste aplicatia magazinului ?(da-40%, nu-60%)
dispozitiv	Categorica	Ce dispozitiv foloseste ? (mobil-50%, tableta-20%, desktop-30%)
timp	Numerica	Cate minute petrece pe magazin? (intre 5 si 59)
frecventa_lunara	Numerica	Cat de des comanda/intra pe magazin ? (intre 0 si 10 ori)
varsta	Numerica	Ce varsta are clientul ?(intre 18 si 59 ani)
metoda_de_plata	Categorica	Foloseste cupoane de reducere ?(da-60%, nu-40%)
recenzii	Numerica	Cate recenzii scrie un client? (intre 0 si 10)

-> Coloana tinta => tip_client

- Fiecarui client i s-a atribuit un **scor compozit** bazat pe toate cele 9 caracteristici. Scorul a fost calculat printr-un sistem de punctaj, in care valorile mari reflecta un client mai valoros (ex: comenzile mai mari, frecventa ridicata, varsta mai mare etc.).
- Dupa calcularea scorurilor, clientii i-am etichetat pe baza percentilei:

Interval **scor**	Eticheta **tip_client**
<= 20%	Low
20-40%	Standard
40-60%	Moderate
60-80%	Vip
> 80%	Ultra Vip

-> Scopul setului:

– Setul este utilizat pentru:

- Antrenarea si evaluarea unui model de clasificare (tip_client)

- Explorarea relatiilor dintre comportamentele clientilor si eticheta atribuita
- Vizualizari si analiza statistica (ex.: distributii, outlieri, corelatii, heatmap-uri, histograme etc.)

6. Analiza exploratorie a datelor (EDA complex)

a) Analiza valorilor lipsa

-> Datorita faptului ca am generat setul de date sintetic, este logic sa nu am nicio valoare lipsa (train.csv si test.csv nu au valori lipsa)

b) Statistici descriptive

1. train.csv

- Ce observam ?
  - Media valorii comenzilor este ~392, iar valorile variaza intre 10 si 1000.
  - Timpul petrecut pe site are o medie ~31 minute.
  - Varsta medie a clientilor este ~39 ani, cuprinsa intre 18 si 59.
  - Clientii plaseaza in medie 5 comenzi pe luna.
  - Cele mai multe comenzi sunt facute de pe mobil.
- Ce suspiciuni/idei putem formula?
  - Valorile pentru comenzile medii sunt foarte variate – pot exista clienti cu comportamente diferite.
  - Multi clienti nu folosesc aplicatia, dar totusi comanda des.
  - Majoritatea prefera plata cu cardul si folosesc cupoane.
- Ce preprocesari ar trebui sa aplicam?
  - Pot verifica daca exista valori aberante (outlieri) pentru preturi sau timp.
  - Variabilele categorice (ex: dispozitiv, metoda de plata) vor trebui transformate in numere inainte de a antrena modelul.

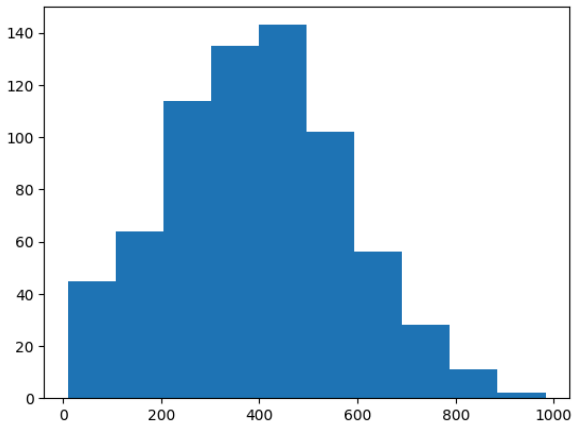
2. test.csv

- Ce observam ?
  - Media valorii comenzilor este ~402, usor mai mare decat in setul de antrenare.
  - Timpul petrecut pe site este similar (~30 minute)..
  - Varsta medie a clientilor este ~38 ani, deci consistenta cu train.csv.
  - Clientii plaseaza in medie 5 comenzi pe luna, cu distributie apropiata de train.
  - Dispozitivul preferat ramane mobil, iar majoritatea folosesc cardul si cupoane.
- Ce suspiciuni/idei putem formula?
  - Distributia e foarte asemanatoare cu setul de antrenare – datele sunt bine separate.
  - Frecventa mare a utilizatorilor care folosesc cupoane sau card poate influenta tip_client.
  - Posibile valori extreme la valoare_comanda_medie, avand max = 919.
- Ce preprocesari ar trebui sa aplicam?
  - Verificarea valorilor aberante (outlieri), mai ales pentru pret si timp..
  - Transformarea valorilor categorice in numerice pentru a putea face predictii.

c) Analiza distributiei variabilelor

1. train.csv

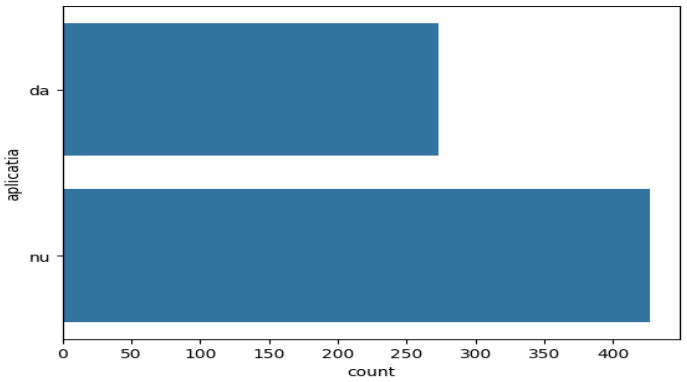
1.1 **`Valoarea medie a unei comenzi (valoare numerica)`**



- Ce observam ?
  - Distributia este concentrata intre 200 si 600 lei, cu un varf in jurul valorii 500.
- Ce suspiciuni/idei putem formula?
  - Valorile mari peste 800 par a fi iesite din tipar si pot influenta negativ modelele.
- Ce preprocesari ar trebui sa aplicam?
  - identificarea si eliminarea outlierilor inainte de antrenare.

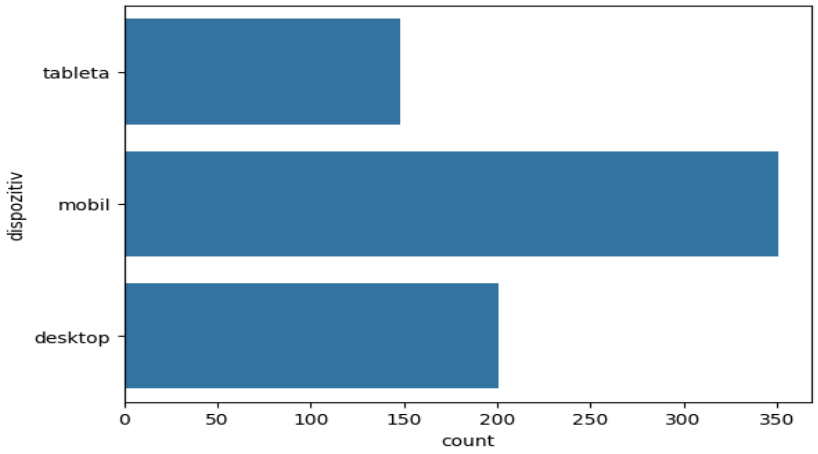
1.2 **`Foloseste aplicatia? (valoare categorica)`**

- Ce observam ?



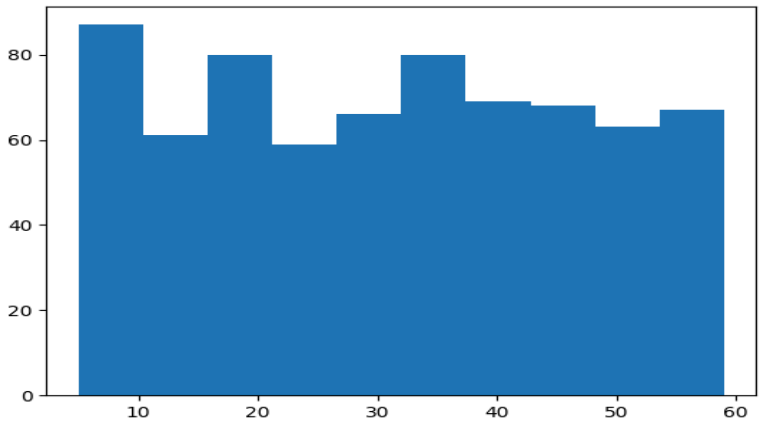
  - Majoritatea clientilor nu folosesc aplicatia, doar o parte mai mica o utilizeaza.
- Ce suspiciuni/idei putem formula?
  - Este posibil ca folosirea aplicatiei sa influenteze pozitiv comportamentul de cumparare.
- Ce preprocesari ar trebui sa aplicam?
  - Variabila trebuie transformata in format numeric binar (ex: 0 = nu, 1 = da).

1.3 `Ce dispozitiv foloseste? (valoare categorica)`



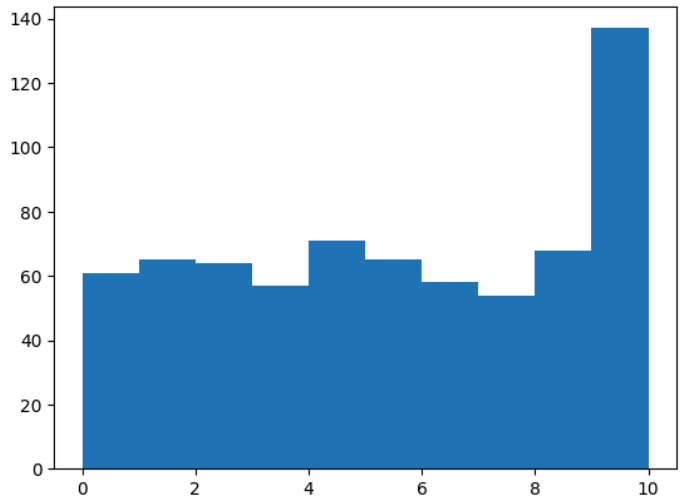
- Ce observam ?
  - Majoritatea comenzilor sunt plasate de pe mobil, urmate de desktop, apoi tableta.
- Ce suspiciuni/idei putem formula?
  - Este posibil ca utilizatorii de desktop sa aiba un comportament de cumparare diferit fata de cei de mobil.
- Ce preprocesari ar trebui sa aplicam?
  - Variabila trebuie transformata in variabile numerice (ex: one-hot encoding sau label encoding).

1.4 `Cat timp petrece pe magazin? (valoare numerica)`



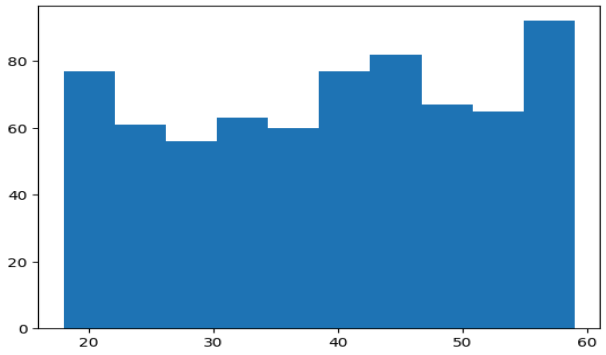
- Ce observam ?
  - Timpul petrecut variaza uniform intre 5 si 60 minute, fara varfuri evidente.
- Ce suspiciuni/idei putem formula?
  - Este posibil ca timpul sa nu fie un factor decisiv in comportamentul clientului.
- Ce preprocesari ar trebui sa aplicam?
  - Putem standardiza valorile pentru o mai buna integrare in model.

1.5 **`Frecventa lunara (valoare numerica)`**



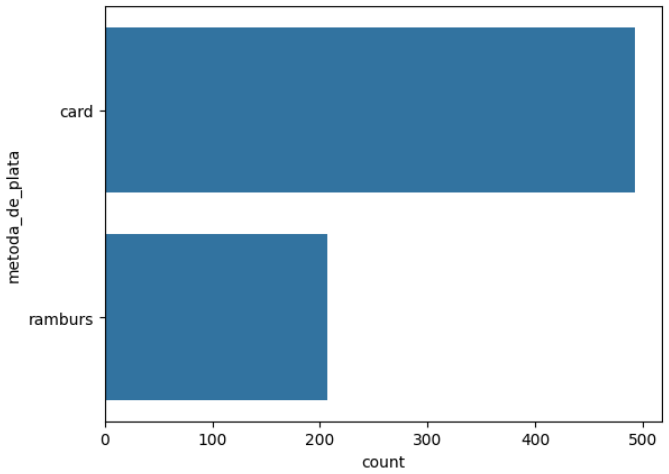
- Ce observam ?
  - Majoritatea clientilor plaseaza intre 0 si 10 comenzi lunar, cu un varf clar la valoarea 10.
- Ce suspiciuni/idei putem formula?
  - Unii clienti comanda foarte des in fiecare luna, ajungand chiar la limita maxima.
- Ce preprocesari ar trebui sa aplicam?
  - Putem verifica daca valoarea 10 apare prea des si daca este nevoie sa o ajustam sau sa o analizam separat.

1.6 **`Varsta (valoare numerica)`**



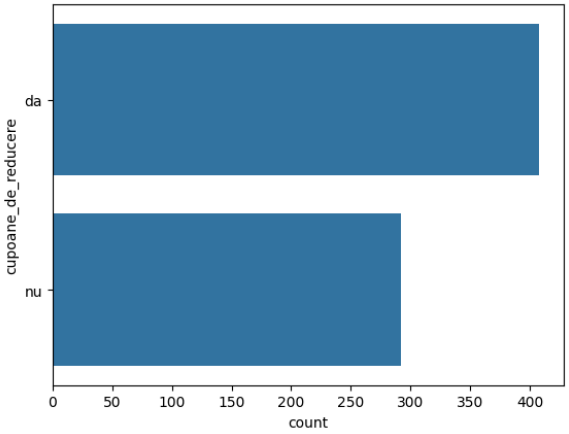
- Ce observam ?
  - Varstele clientilor sunt distribuite destul de uniform intre 18 si 59 de ani, cu usoare cresteri spre 40+.
- Ce suspiciuni/idei putem formula?
  - Este posibil ca magazinul sa fie popular in randul adultilor de varsta medie.
- Ce preprocesari ar trebui sa aplicam?
  - Putem imparti varstele pe grupe (ex: <30, 30–45, >45) pentru o analiza mai clara a comportamentului pe segmente de varsta.

1.7 **Metoda de plata (valoare categorica)**



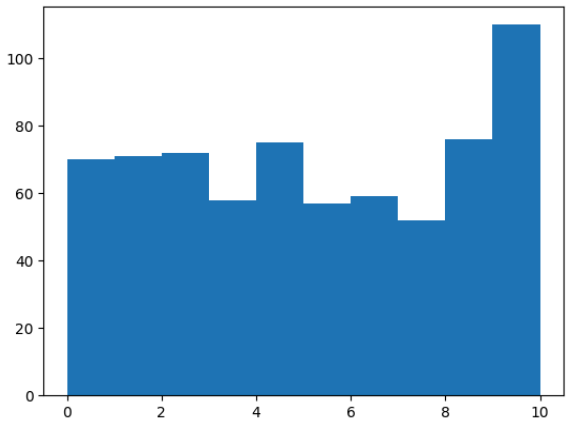
- Ce observam ?
  - Majoritatea clientilor prefera sa plateasca cu cardul, in timp ce o parte mai mica alege rambursul..
- Ce suspiciuni/idei putem formula?
  - Plata cu cardul ar putea fi asociata cu clienti mai activi sau fideli magazinului.
- Ce preprocesari ar trebui sa aplicam?
  - Transformam aceasta variabila categorica in numerica (ex: 0 pentru ramburs, 1 pentru card) -> Label Encoding.

1.8 **Foloseste cupoane de reducere? (valoare categorica)**



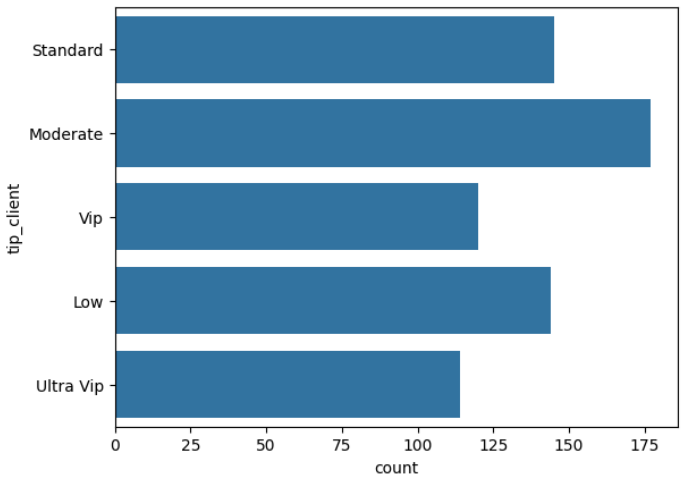
- Ce observam ?
  - Mai multi clienti folosesc cupoane de reducere decat cei care nu le folosesc.
- Ce suspiciuni/idei putem formula?
  - Folosirea cupoanelor poate indica o preocupare crescuta pentru reduceri sau fidelitate fata de magazin
- Ce preprocesari ar trebui sa aplicam?
  - Transformam aceasta variabila categorica in numerica (ex: 0 pentru "nu", 1 pentru "da") -> Label Encoding.

1.9 **`Cate recenzii acorda? (valoare numerica)`**



- Ce observam ?
  - Cei mai multi clienti ofera intre 0 si 10 recenzii, cu o usoara crestere spre capatul superior.
- Ce suspiciuni/idei putem formula?
  - Unii clienti sunt foarte activi in a oferi feedback, ceea ce poate semnala un interes mai mare pentru magazin.
- Ce preprocesari ar trebui sa aplicam?
  - Putem scala aceasta variabila sau o putem pastra ca atare daca modelul o gestioneaza bine.

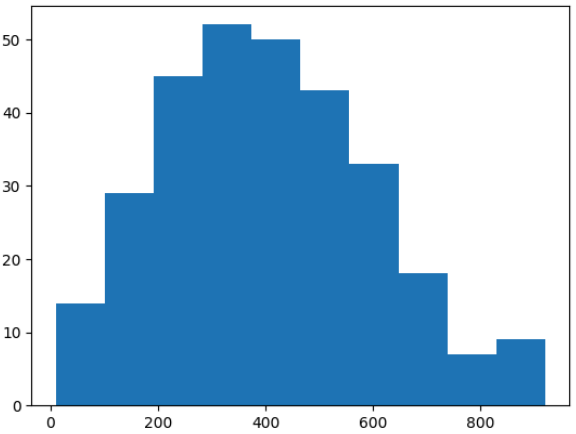
1.10 **`Ce tip de client este? (valoare categorica) -> `*`Coloana tinta`**



- Ce observam ?
  - Categoria „Moderate” are cei mai multi clienti, iar „Vip” si „Ultra Vip” sunt mai putin frecvente.
- Ce suspiciuni/idei putem formula?
  - Distributia nu este echilibrata, ceea ce poate afecta performanta modelelor de clasificare.
- Ce preprocesari ar trebui sa aplicam?
  - Coloana tip_client nu trebuie transformata, deoarece este tinta problemei si poate fi folosita ca atare in majoritatea modelelor de clasifica

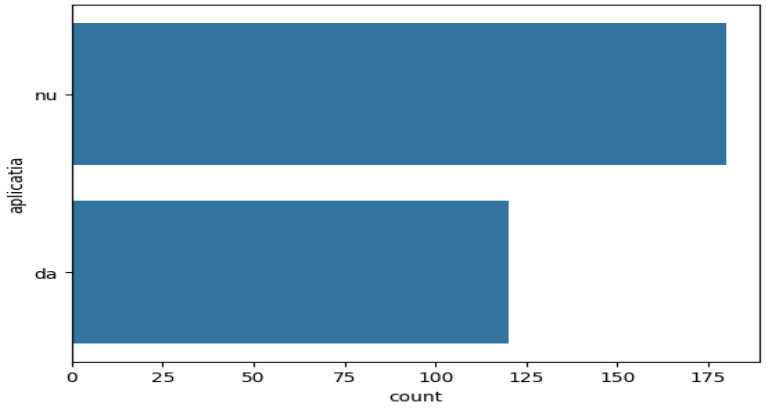
2. test.csv

2.1 **`Valoarea medie a unei comenzi (valoare numerica)`**



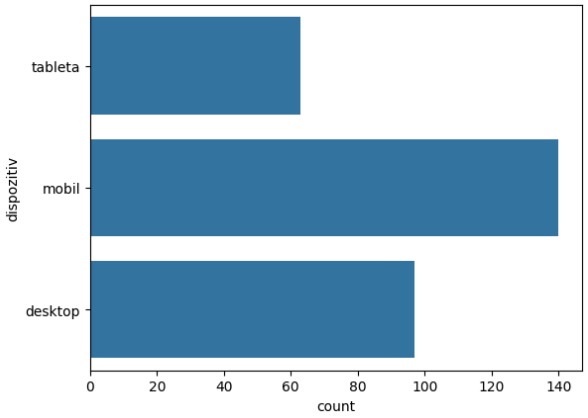
- Ce observam ?
  - Majoritatea comenzilor au o valoare medie intre 200 si 500, cu o distributie usor asimetrica spre dreapta.
- Ce suspiciuni/idei putem formula?
  - Exista cativa clienti care plaseaza comenzi semnificativ mai mari, posibil clienti premium.
- Ce preprocesari ar trebui sa aplicam?
  - Se pot verifica valorile mari pentru outlieri si aplica o scalare pentru a aduce valorile pe un interval comparabil.

2.2 **`Foloseste aplicatia? (valoare categorica)`**



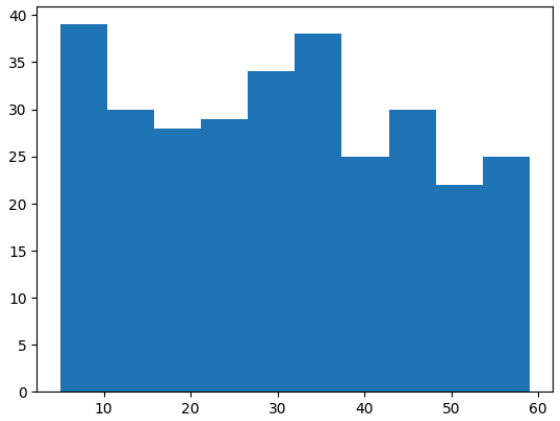
- Ce observam ?
  - Majoritatea utilizatorilor din setul de test nu folosesc aplicatia magazinului.
- Ce suspiciuni/idei putem formula?
  - Unii clienti prefera sa nu foloseasca aplicatia, posibil din lipsa de interes sau din motive de confort.
- Ce preprocesari ar trebui sa aplicam?
  - Trebuie convertita variabila in format numeric (ex. 1 pentru "da", 0 pentru "nu") pentru antrenarea modelului -> Label Encoding.

2.3 `Ce dispozitiv foloseste? (valoare categorica)`



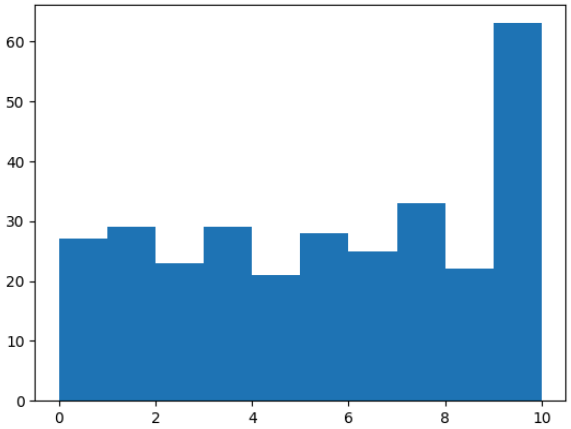
- Ce observam ?
  - Majoritatea utilizatorilor acceseaza platforma de pe mobil, urmati de cei de pe desktop si tablete.
- Ce suspiciuni/idei putem formula?
  - Mobilul pare sa fie cel mai comod mijloc de acces, ceea ce poate influenta comportamentele de cumparare rapide.
- Ce preprocesari ar trebui sa aplicam?
  - Vom transforma aceste valori categorice (mobil, desktop, tableta) in valori numerice pentru a le putea introduce in model.

2.4 `Cat timp petrece pe magazin? (valoare numerica)`



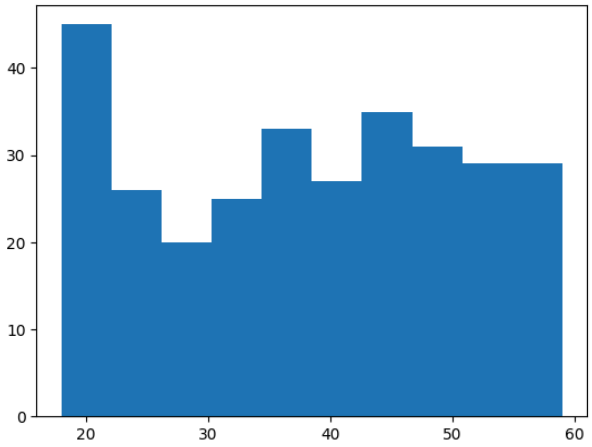
- Ce observam ?
  - Majoritatea clientilor petrec intre 5 si 35 de minute pe site, dar sunt si unii care stau aproape o ora.
- Ce suspiciuni/idei putem formula?
  - Clientii care petrec mai mult timp pe platforma ar putea analiza mai atent ofertele sau produsele.
- Ce preprocesari ar trebui sa aplicam?
  - Putem verifica existenta valorilor extreme si normaliza coloana pentru o interpretare mai usoara de catre modele.

2.5 `Frecventa lunara (valoare numerica)`



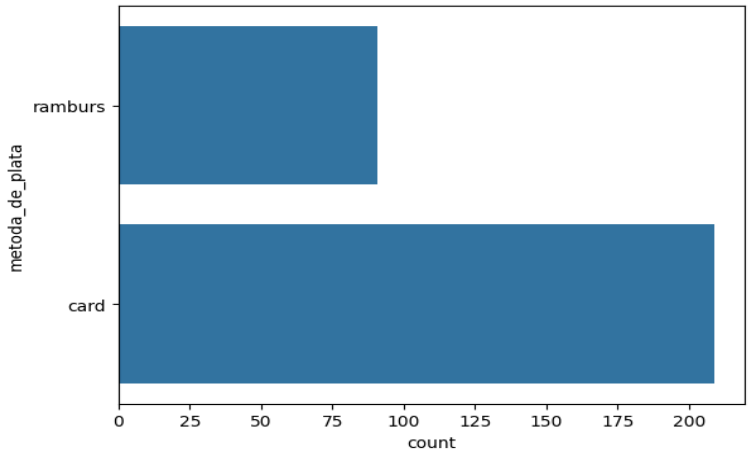
- Ce observam ?
  - Cei mai multi clienti plaseaza aproximativ 10 comenzi pe luna, ceea ce indica utilizatori activi.
- Ce suspiciuni/idei putem formula?
  - Ar putea exista un segment fidel de clienti care comanda lunar aproape saptamanal.
- Ce preprocesari ar trebui sa aplicam?
  - Putem normaliza aceasta coloana pentru a o aduce intr-un interval comparabil cu alte variabile numerice.

2.6 `Varsta (valoare numerica)`



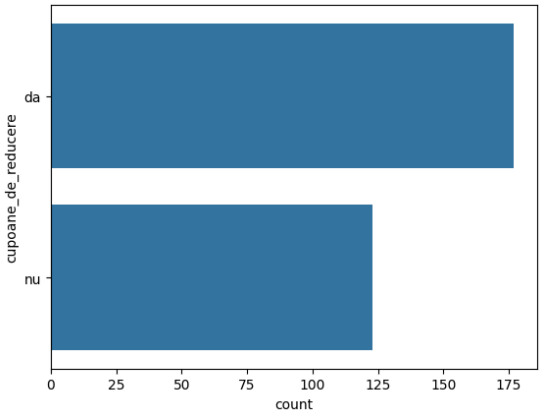
- Ce observam ?
  - Majoritatea clientilor au varste intre 18 si 25 de ani, cu o distributie destul de uniforma dupa acest interval.
- Ce suspiciuni/idei putem formula?
  - Publicul pare sa fie in principal tanar, ceea ce poate influenta comportamentul de cumparare.
- Ce preprocesari ar trebui sa aplicam?
  - Putem normaliza valorile varstei pentru a fi mai usor de integrat in algoritmi de clasificare (Logistic).

2.7 **Metoda de plata (valoare categorica)**



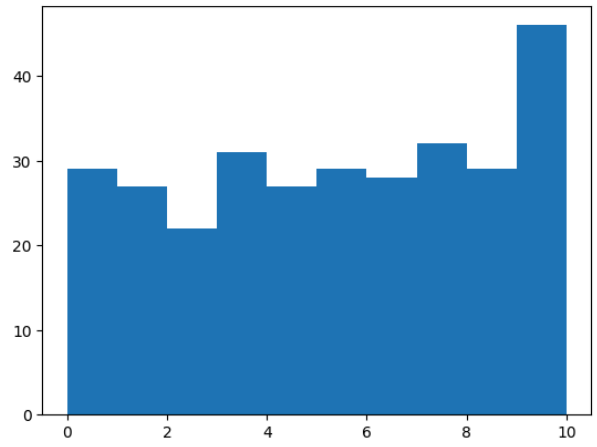
- Ce observam ?
  - Majoritatea clientilor prefera sa plateasca cu cardul, in timp ce o parte mai mica aleg rambursul.
- Ce suspiciuni/idei putem formula?
  - Clientii care platesc cu cardul pot fi mai increzatori in platforma sau mai obisnuiti cu comenzile online.
- Ce preprocesari ar trebui sa aplicam?
  - Transformarea metodelor de plata in valori numerice (ex: card = 1, ramburs = 0) pentru a putea fi folosite in modele -> Label Encoding.

2.8 **Foloseste cupoane de reducere? (valoare categorica)**



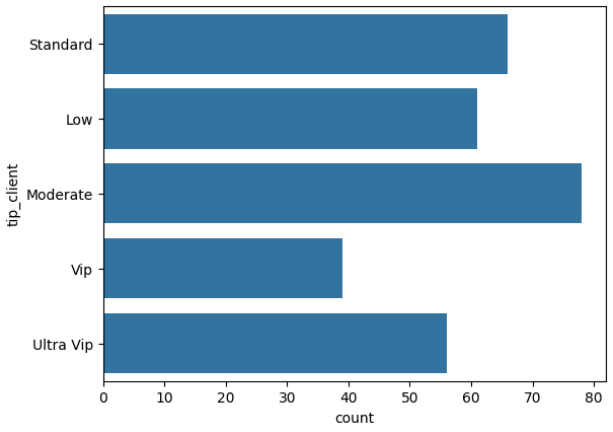
- Ce observam ?
  - Mai multi clienti folosesc cupoane de reducere decat cei care nu le folosesc.
- Ce suspiciuni/idei putem formula?
  - Este posibil ca utilizarea cupoanelor sa indice clienti mai activi, care cauta oferte sau reduceri frecvent.
- Ce preprocesari ar trebui sa aplicam?
  - Variabila trebuie codificata numeric pentru a putea fi folosita in modele.

2.9 **`Cate recenzii acorda? (valoare numerica)`**



- Ce observam ?
  - Majoritatea clientilor ofera intre 0 si 10 recenzii, cu o usoara crestere spre valoarea maxima.
- Ce suspiciuni/idei putem formula?
  - Clientii care scriu mai multe recenzii ar putea fi mai implicati sau mai multumiti.
- Ce preprocesari ar trebui sa aplicam?
  - Putem verifica daca acesta variabila influenteaza tipul de client si o putem pastra ca numerica.

2.10 **`Ce tip de client este? (valoare categorica) ->`** ***`Coloana tinta`***

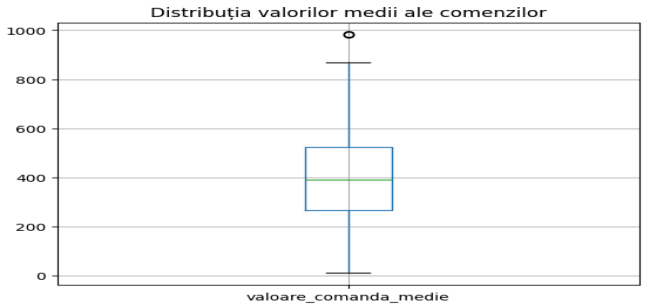


- Ce observam ?
  - Cele mai multe cazuri sunt Moderate, urmate de Standard si Low, iar cele mai putine sunt Vip.
- Ce suspiciuni/idei putem formula?
  - Distributia e relativ echilibrata, dar clientii Vip sunt mai rari, deci poate fi o clasa dificil de prezis.
- Ce preprocesari ar trebui sa aplicam?
  - Coloana tip_client nu trebuie transformata, deoarece este tinta problemei si poate fi folosita ca atare in majoritatea modelelor de clasificare.

d) Detectarea outlierilor

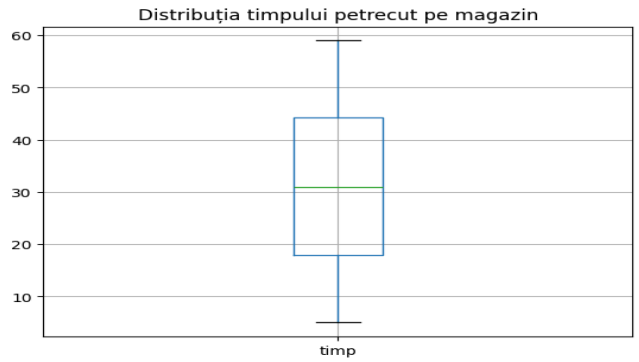
1. train.csv

1.1 **`Valoarea medie a unei comezi`**



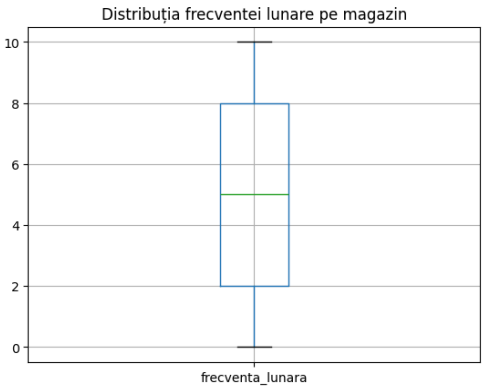
- Ce observam ?
  - Majoritatea valorilor se afla intre 300 si 500, iar mediana este aproape de 400.
- Ce suspiciuni/idei putem formula?
  - Exista o valoare aberanta in jur de 990, care poate fi un client care plaseaza comenzi foarte scumpe.
- Ce preprocesari ar trebui sa aplicam?
  - Putem analiza daca outlierul este justificat sau daca ar trebui eliminat pentru a nu influenta antrenarea modelului.

1.2 **`Timpul petrecut pe magazin`**



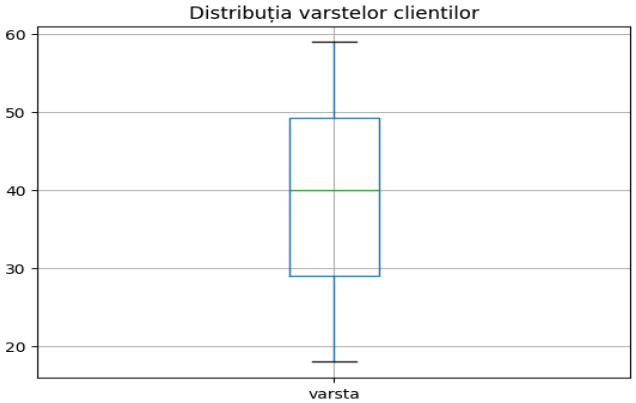
- Ce observam ?
  - Majoritatea valorilor se afla intre 19 si 45 de minute, cu o mediana de aproximativ 32.
- Ce suspiciuni/idei putem formula?
  - Distributia este destul de echilibrata si nu apar valori aberante.
- Ce preprocesari ar trebui sa aplicam?
  - Nu este necesara eliminarea de outlieri, dar putem normaliza valorile daca modelul folosit o cere.

1.3 `Frecventa lunara`



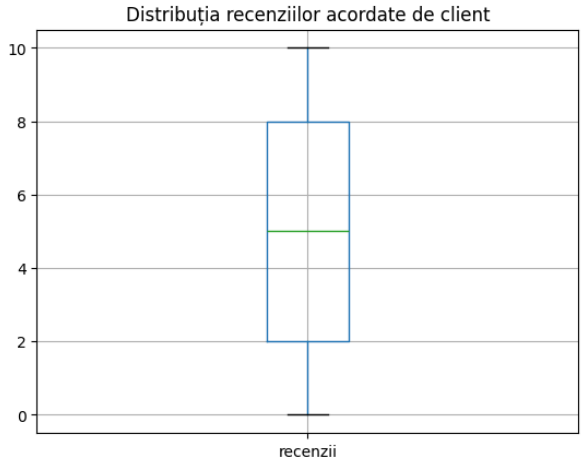
- Ce observam ?
  - Majoritatea valorilor sunt intre 2 si 8 comenzi pe luna, cu o mediana in jur de 5.
- Ce suspiciuni/idei putem formula?
  - Valorile sunt bine distribuite si nu apar outlieri.
- Ce preprocesari ar trebui sa aplicam?
  - Nu este necesara eliminarea valorilor, dar coloana poate fi standardizata daca modelul o impune.

1.4 `Varsta clientului`



- Ce observam ?
  - Majoritatea clientilor au varste cuprinse intre 29 si 49 de ani, iar mediana este 40.
- Ce suspiciuni/idei putem formula?
  - Distributia este echilibrata si nu apar valori aberante..
- Ce preprocesari ar trebui sa aplicam?
  - Coloana poate fi pastrata ca atare sau standardizata in functie de model.

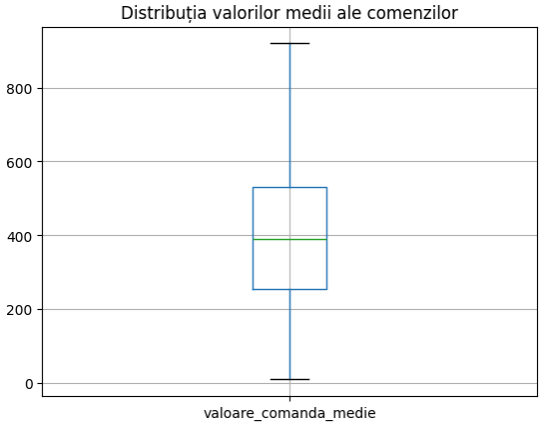
1.5 **`Recenzii`**



- Ce observam ?
  - Majoritatea clientilor ofera intre 2 si 8 recenzii, iar mediana este in jur de 5.
- Ce suspiciuni/idei putem formula?
  - Exista clienti care ofera foarte putine sau foarte multe recenzii, dar nu apar outlieri evidenti.
- Ce preprocesari ar trebui sa aplicam?
  - Coloana poate fi normalizata sau folosita direct, fiind deja scalata intre 0 si 10.

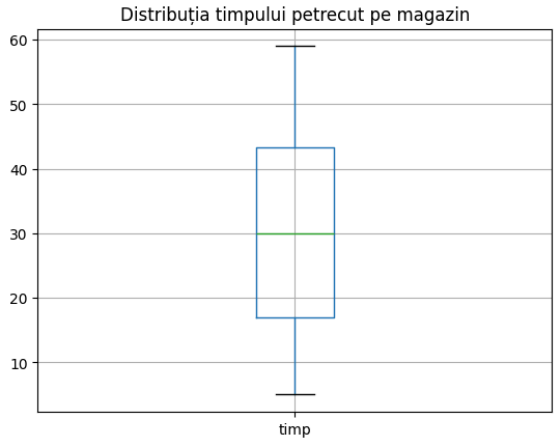
2. test.csv

2.1 **`Valoarea medie a unei comezi`**



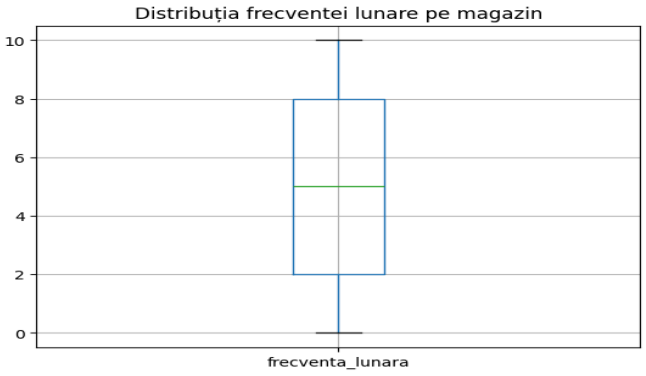
- Ce observam ?
  - Majoritatea valorilor sunt intre 250 si 550, iar mediana este aproape de 400.
- Ce suspiciuni/idei putem formula?
  - Nu exista outlieri evidenti, dar intervalul este destul de larg, ceea ce indica variatii mari intre clienti.
- Ce preprocesari ar trebui sa aplicam?
  - Putem normaliza coloana pentru a reduce impactul valorilor extreme in antrenarea modelului.

2.2 `Timpul petrecut pe magazin`



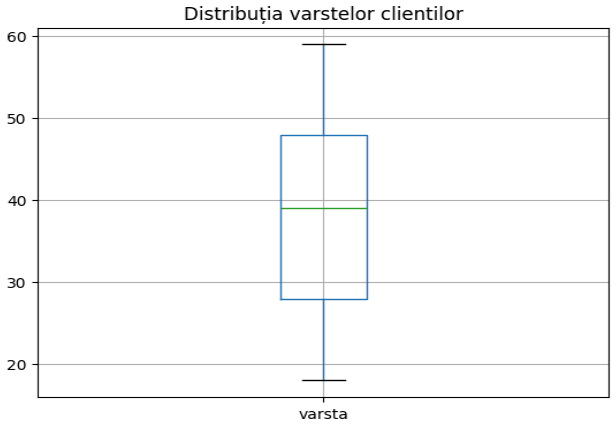
- Ce observam ?
  - Majoritatea clientilor petrec intre 18 si 44 de minute in magazin, iar mediana este 30.
- Ce suspiciuni/idei putem formula?
  - Distributia este echilibrata, fara valori aberante vizibile..
- Ce preprocesari ar trebui sa aplicam?
  - Putem scala valorile pentru a fi comparabile cu alte caracteristici numerice.

2.3 `Frecventa lunara`



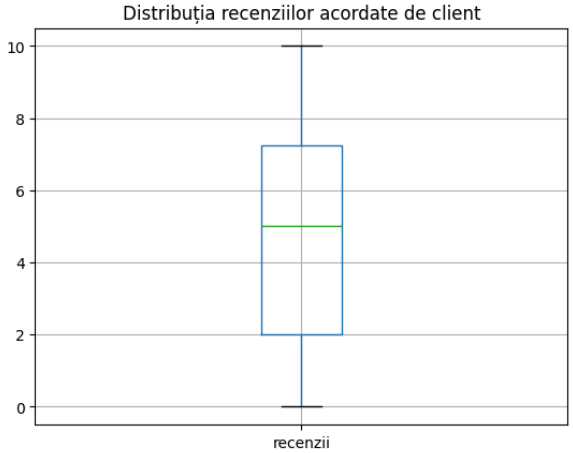
- Ce observam ?
  - Majoritatea clientilor plaseaza intre 2 si 8 comenzi pe luna, cu o mediana in jur de 5.
- Ce suspiciuni/idei putem formula?
  - Distributia este simetrica si nu exista valori aberante.
- Ce preprocesari ar trebui sa aplicam?
  - Valorile pot fi scalate daca sunt folosite in modele sensibile la intervale.

2.4 **`Varsta clientului`**



- Ce observam ?
  - Majoritatea clientilor au varsta intre 28 si 48 de ani, cu o mediana in jur de 39.
- Ce suspiciuni/idei putem formula?
  - Distributia este echilibrata si nu apar valori extreme sau neasteptate.
- Ce preprocesari ar trebui sa aplicam?
  - Valoarea poate fi normalizata daca modelul folosit este influentat de scara variabilelor.

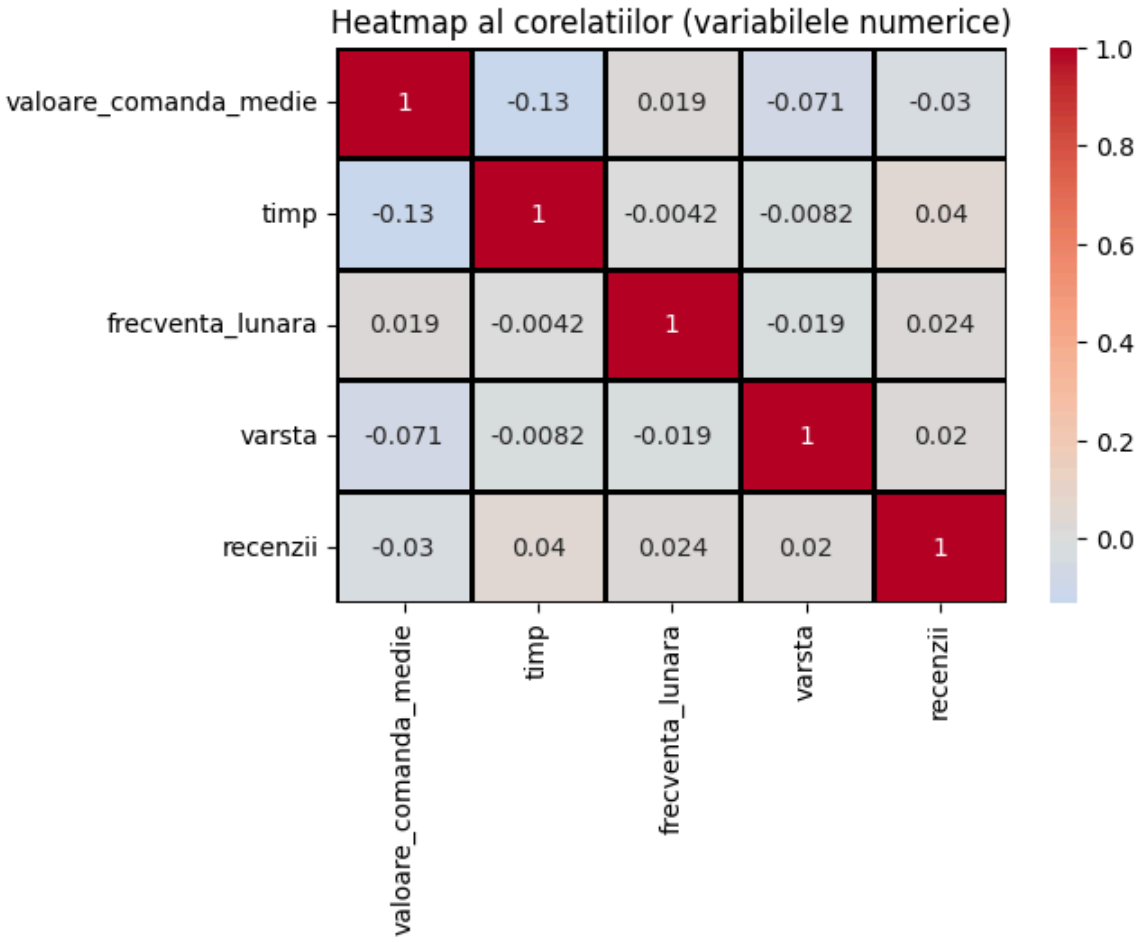
2.5 **`Recenzii`**



- Ce observam ?
  - Majoritatea clientilor ofera intre 2 si 7 recenzii, iar mediana este in jur de 5.
- Ce suspiciuni/idei putem formula?
  - Distributia este echilibrata si nu exista valori aberante evidente.
- Ce preprocesari ar trebui sa aplicam?
  - Putem normaliza valorile daca modelul folosit este sensibil la scara numerica.

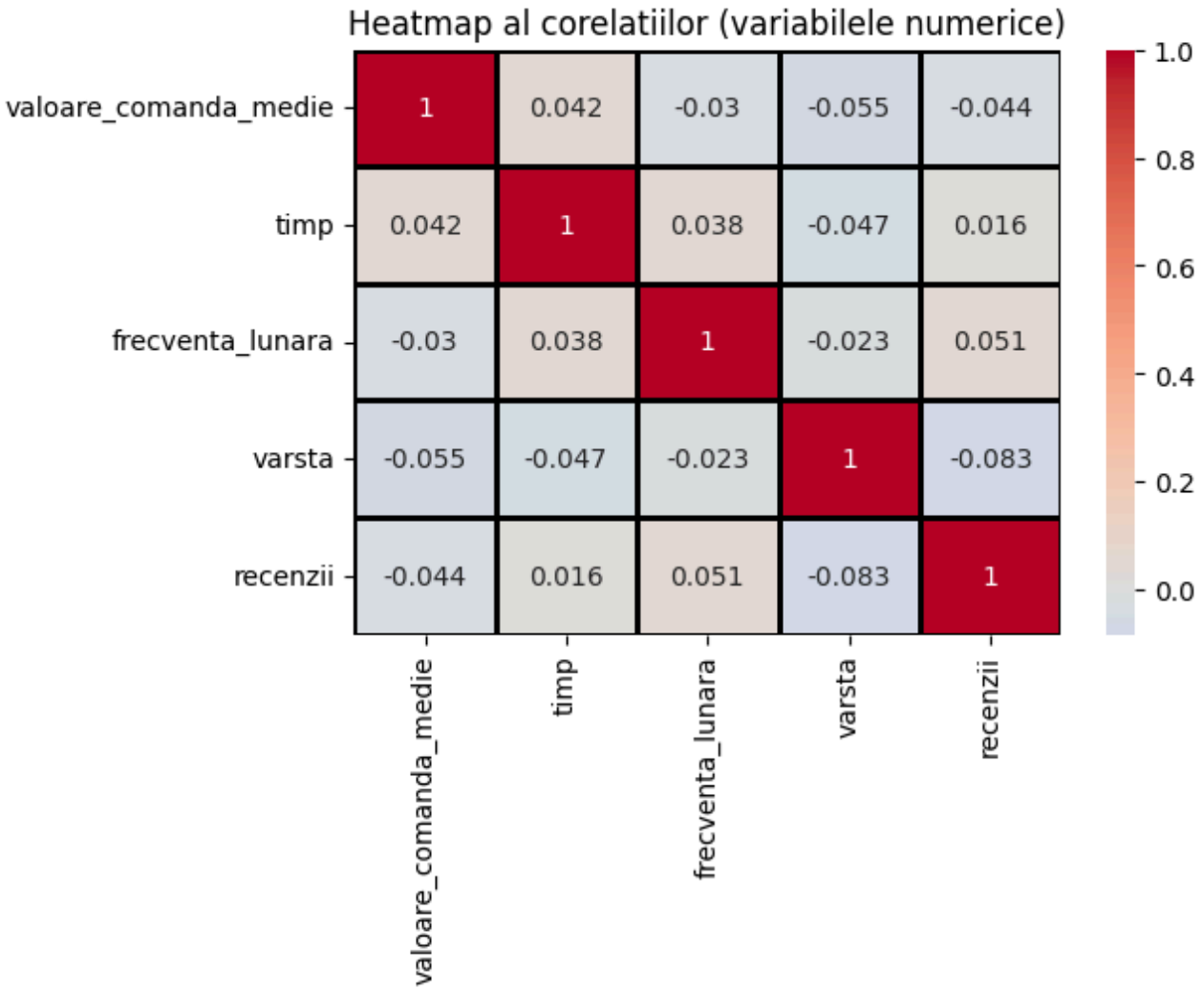
e) Analiza corelatiilor

1. train.csv



- Ce observam ?
  - Corelatiile dintre variabilele numerice sunt foarte slabe, majoritatea avand valori apropiate de 0.
- Ce suspiciuni/idei putem formula?
  - Variabilele par a fi aproape independente intre ele, deci nu exista relatii liniare semnificative intre ele.
- Ce preprocesari ar trebui sa aplicam?
  - Nu este nevoie de eliminare a colinearitatilor, dar putem standardiza valorile daca folosim modele sensibile la scala.

2. test.csv

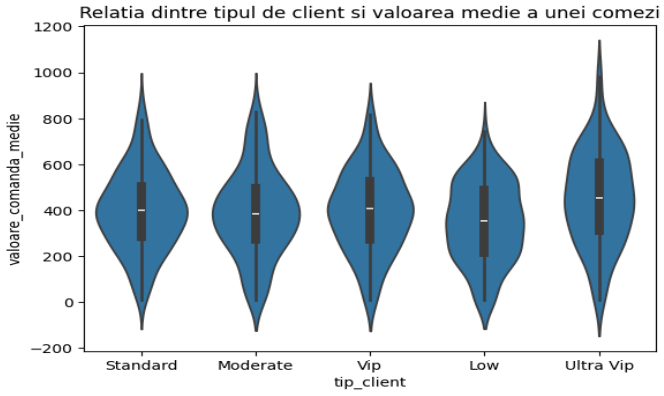


- Ce observam ?
  - Corelatiile dintre variabile sunt foarte apropiate de 0, deci nu exista legaturi liniare evidente.
- Ce suspiciuni/idei putem formula?
  - Variabilele sunt slab corelate intre ele, ceea ce inseamna ca fiecare aduce informatii diferite.
- Ce preprocesari ar trebui sa aplicam?
  - Pot standardiza variabilele pentru modele care tin cont de scala, dar nu este nevoie sa eliminam vreo variabila.

f) Analiza relatiilor cu variabila tinta

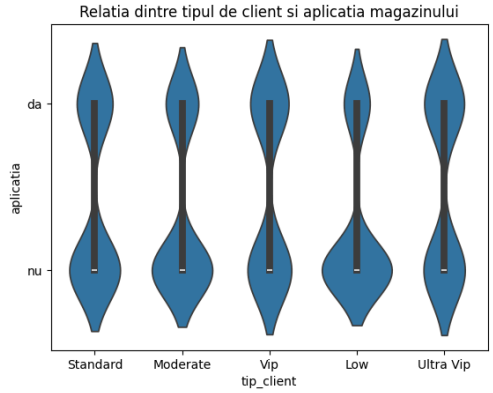
1. train.csv

1.1 **`valoare_comanda_medie <-> tip_client`**



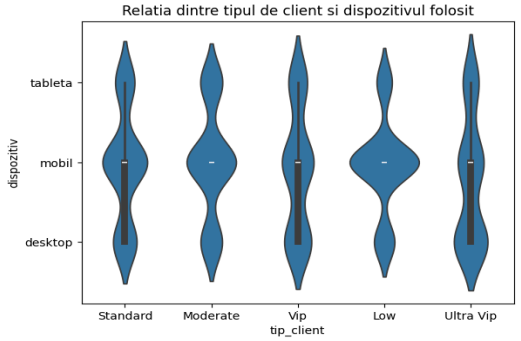
- Ce observam ?
  - Clientii din categoriile Ultra Vip si Vip tind sa aiba comenzi cu valori mai mari decat cei din categoriile Low si Standard.
- Ce suspiciuni/idei putem formula?
  - Exista o legatura pozitiva intre tipul de client si valoarea medie a unei comenzi, adica clientii mai valorosi cheltuie mai mult.
- Ce preprocesari ar trebui sa aplicam?
  - Coloana valoare_comanda_medie trebuie scalata pentru ca are valori mari si modele pot fi afectate.

1.2 **`aplicatia <-> tip_client`**



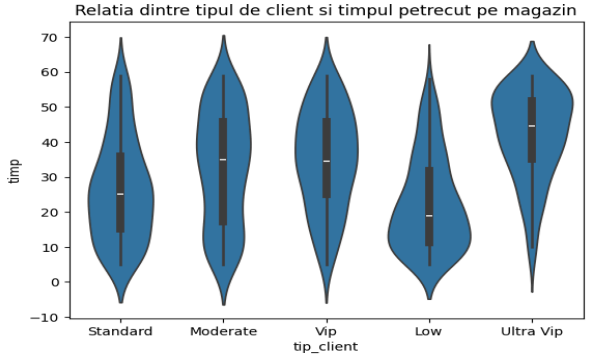
- Ce observam ?
  - Majoritatea clientilor, indiferent de tip, folosesc aplicatia magazinului. Proportiile sunt similare pe toate categoriile.
- Ce suspiciuni/idei putem formula?
  - Utilizarea aplicatiei nu pare sa fie un factor distinctiv intre tipurile de clienti, dar ar putea contribui indirect prin cresterea interactiunii.
- Ce preprocesari ar trebui sa aplicam?
  - Coloana aplicatia trebuie transformata in variabila numerica (ex: 0 si 1).

1.3 `dispozitiv <-> tip_client`



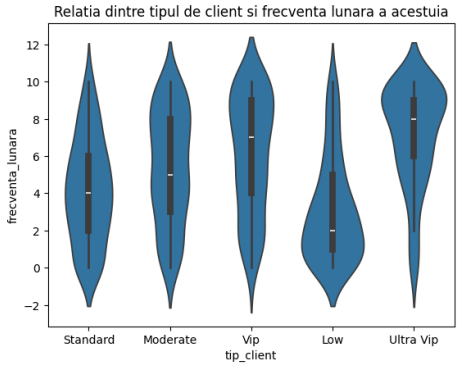
- Ce observam ?
  - Dispozitivul mobil este cel mai folosit pentru toate tipurile de clienti. Desktop-ul are o usoara prezenta la clientii Vip si Standard. Tableta este folosita marginal.
- Ce suspiciuni/idei putem formula?
  - Clientii prefera mobilul indiferent de scorul lor, ceea ce sugereaza ca experienta mobila este esentiala. Tableta are o utilizare redusa, posibil din cauza interfetei sau obisnuintei.
- Ce preprocesari ar trebui sa aplicam?
  - Aplicam one-hot encoding pentru variabila dispozitiv.

1.4 `timp <-> tip_client`



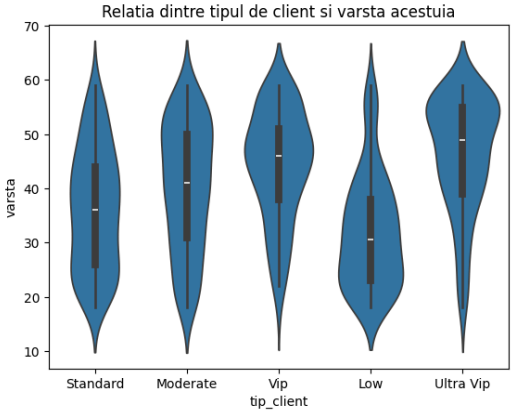
- Ce observam ?
  - Clientii Ultra Vip petrec cel mai mult timp in magazin, urmati de cei Vip si Moderate. Clientii Low petrec cel mai putin timp.
- Ce suspiciuni/idei putem formula?
  - Clientii Ultra Vip petrec cel mai mult timp in magazin, urmati de cei Vip si Moderate. Clientii Low petrec cel mai putin timp.
- Ce preprocesari ar trebui sa aplicam?
  - Standardizare a variabilei timp si verificare pentru outlieri.

1.5 **`frecventa_lunara <-> tip_client`**



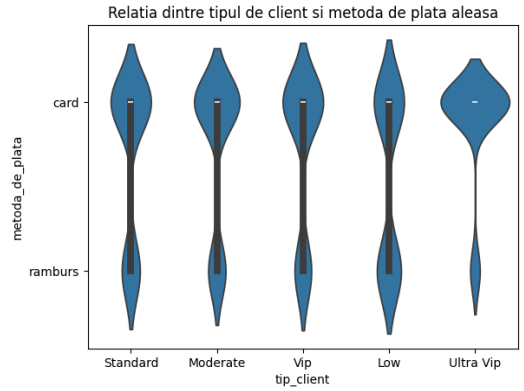
- Ce observam ?
  - Clientii Ultra Vip si Vip au frecvente lunare ridicate, concentrate in jurul valorilor 8–10. Clientii Low au frecvente scazute, majoritatea sub 3.
- Ce suspiciuni/idei putem formula?
  - Exista o corelatie pozitiva intre frecventa lunara si tipul de client. Frecventa ridicata poate contribui direct la obtinerea unui scor mai mare.
- Ce preprocesari ar trebui sa aplicam?
  - Putem standardiza frecventa_lunara pentru modele sensibile la scala. De asemenea, putem verifica distributia pentru outlieri.

1.6 **`varsta <-> tip_client`**



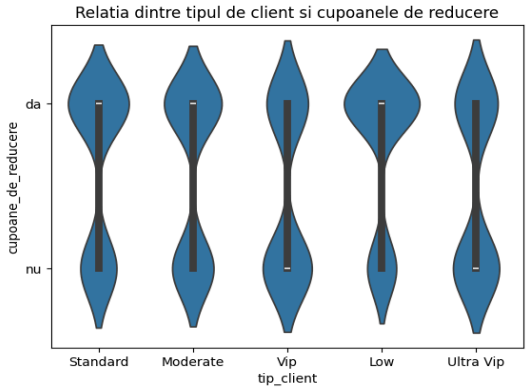
- Ce observam ?
  - Clientii Ultra Vip si Vip au in general varste mai mari, cu mediane peste 45 de ani. Clientii Low au varste mai mici, in medie sub 35 de ani.
- Ce suspiciuni/idei putem formula?
  - Scorul clientului pare sa creasca odata cu varsta. Posibil ca experienta si veniturile mai mari sa contribuie la loialitate si clasificare superioara.
- Ce preprocesari ar trebui sa aplicam?
  - Normalizare sau standardizare a varstei, daca este folosita ca predictor. Eventual discretizare in intervale daca folosim modele bazate pe categorii.

1.7 `metoda_de_plata <-> tip_client`



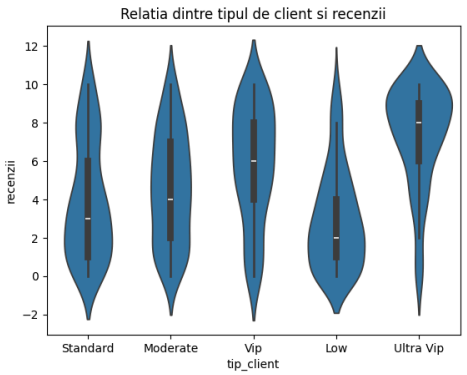
- Ce observam ?
  - Clientii din toate categoriile prefera in majoritate plata cu cardul. Pentru Ultra Vip, plata cu cardul este aproape exclusiv utilizata.
- Ce suspiciuni/idei putem formula?
  - Clientii cu scor mai mare (Ultra Vip) prefera metode moderne de plata, ceea ce poate sugera un profil mai digitalizat sau mai activ online.
- Ce preprocesari ar trebui sa aplicam?
  - Transformam metoda_de_plata in variabila numerica (ex: 0 pentru ramburs, 1 pentru card). Putem folosi one-hot encoding daca adaugam alte metode in viitor.

1.8 `cupoane_de_reducere <-> tip_client`



- Ce observam ?
  - Proportia celor care folosesc cupoane este vizibil mai mare decat a celor care nu folosesc, indiferent de tipul de client.
- Ce suspiciuni/idei putem formula?
  - Folosirea cupoanelor este un comportament comun in toate categoriile, posibil pentru ca aplicatia incurajeaza acest lucru sau pentru ca ofertele sunt atractive. Tipul de client influenteaza mai putin decizia de a folosi cuponul.
- Ce preprocesari ar trebui sa aplicam?
  - Convertim cupoane_de_reducere in binar (0/1). Putem verifica echilibrul de clase si aplicam stratificare la split, daca este cazul.

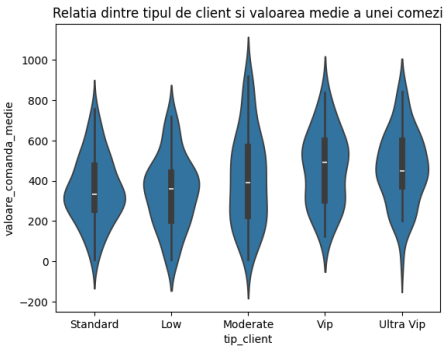
1.9 **`recenzii <-> tip_client`**



- Ce observam ?
  - Clientii Ultra Vip si Vip ofera cele mai multe recenzii. Clientii Low ofera foarte putine, majoritatea avand intre 0 si 2 recenzii.
- Ce suspiciuni/idei putem formula?
  - Numarul de recenzii este un indicator bun al implicarii clientului si pare corelat cu scorul clientului. Cei mai activi sunt si cei mai valorosi.
- Ce preprocesari ar trebui sa aplicam?
  - Verificam distributia pentru outlieri si aplicam normalizare.

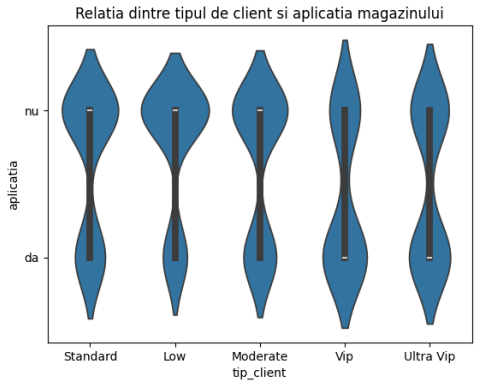
2. test.csv

   2.1 **`valoare_comanda_medie <-> tip_client`**



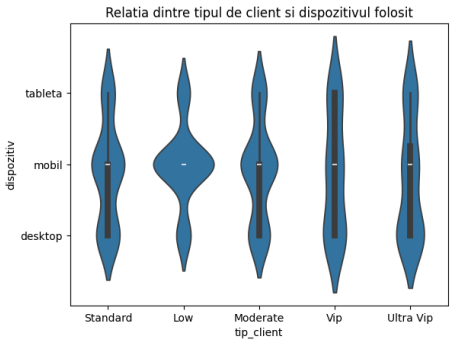
- Ce observam ?
  - Clientii Ultra Vip si Vip au in general comenzi cu valoare medie mai mare. Clientii Low au valori mai scazute si distribuite in intervale mai mici.
- Ce suspiciuni/idei putem formula?
  - Valoarea comenzii este un indicator important al scorului clientului. Clientii mai valorosi tind sa cheltuie mai mult per comanda.
- Ce preprocesari ar trebui sa aplicam?
  - Aplicam standardizare pentru a reduce impactul extremelor. Verificam existenta outlierilor si curatam daca este necesar.

2.2 **aplicatia <-> tip_client**



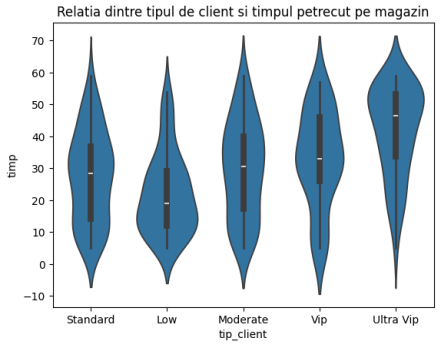
- Ce observam ?
  - Utilizarea aplicatiei este mai frecventa in randul clientilor Vip si Ultra Vip, comparativ cu celelalte categorii.
- Ce suspiciuni/idei putem formula?
  - Aplicatia poate contribui la cresterea implicarii clientului si la clasificarea lui intr-o categorie superioara.
- Ce preprocesari ar trebui sa aplicam?
  - Standardizare sau normalizare a variabilei aplicatia (codificata numeric, de ex. 0/1).

2.3 **dispozitiv <-> tip_client**



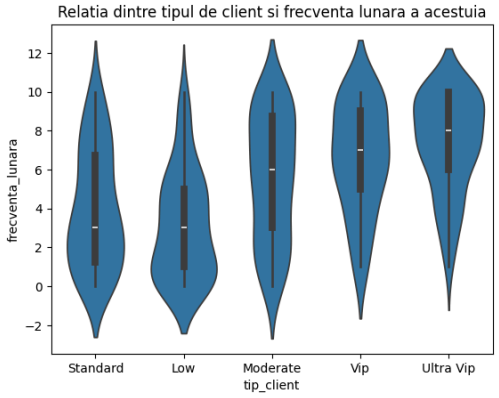
- Ce observam ?
  - Mobilul este cel mai folosit dispozitiv in toate categoriile. Tableta are o crestere vizibila la clientii Vip si Ultra Vip.
- Ce suspiciuni/idei putem formula?
  - Clientii cu scor mare par sa foloseasca mai frecvent tableta, ceea ce poate indica un comportament diferit de navigare sau cumparare.
- Ce preprocesari ar trebui sa aplicam?
  - Standardizare sau normalizare a valorilor codificate ale variabilei dispozitiv.

2.4 `timp <-> tip_client`



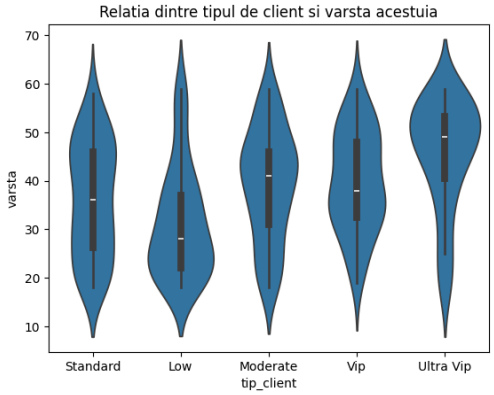
- Ce observam ?
  - Clientii Ultra Vip petrec cel mai mult timp in magazin, urmati de cei Vip si Moderate. Clientii Low au un timp semnificativ mai scazut.
- Ce suspiciuni/idei putem formula?
  - Timpul petrecut in magazin pare sa fie un indicator puternic al tipului de client si al implicarii sale.
- Ce preprocesari ar trebui sa aplicam?
  - Standardizare sau normalizare a variabilei timp.

2.5 `frecventa_lunara <-> tip_client`



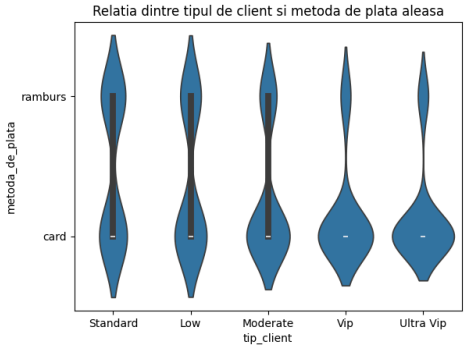
- Ce observam ?
  - Clientii Ultra Vip si Vip au cele mai mari frecvente lunare. Clientii Low au cele mai mici frecvente.
- Ce suspiciuni/idei putem formula?
  - Frecventa lunara este un indicator important al tipului de client si reflecta nivelul de implicare.
- Ce preprocesari ar trebui sa aplicam?
  - Standardizare sau normalizare a variabilei frecventa_lunara.

2.6 **`varsta <-> tip_client`**



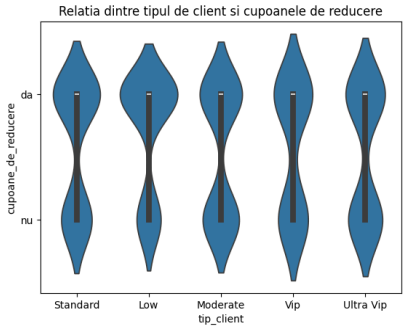
- Ce observam ?
  - Clientii Ultra Vip si Vip au varste mai mari in medie. Clientii Low sunt predominant tineri.
- Ce suspiciuni/idei putem formula?
  - Varsta pare sa fie corelata pozitiv cu tipul de client — clientii mai in varsta tind sa aiba un scor mai mare.
- Ce preprocesari ar trebui sa aplicam?
  - Standardizare sau normalizare a variabilei varsta.

2.7 **`metoda_de_plata <-> tip_client`**



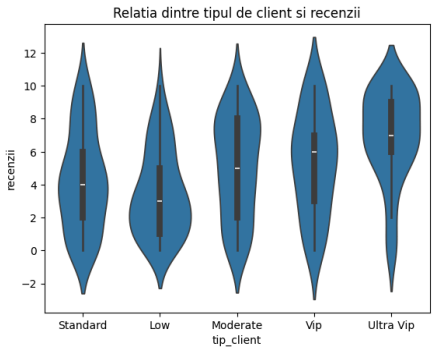
- Ce observam ?
  - Standard, Low si Moderate aleg frecvent plata ramburs, cu o distributie destul de echilibrata intre ramburs si card.
  - Vip si Ultra Vip prefera in mod clar plata cu cardul, aproape exclusiv.
- Ce suspiciuni/idei putem formula?
  - Clientii cu scor mic (Standard, Low, Moderate) pot fi mai traditionali sau reticenti fata de plata online. In schimb, cei cu scor mare (Vip, Ultra Vip) sunt mai familiarizati cu tehnologia si prefera plata rapida prin card..
- Ce preprocesari ar trebui sa aplicam?
  - Standardizare sau normalizare a variabilei metoda_de_plata (codificata numeric).

2.8 **`cupoane_de_reducere <-> tip_client`**



- Ce observam ?
  - Toti clientii, indiferent de tip, folosesc cupoane de reducere intr-o proportie semnificativ mai mare decat cei care nu folosesc.
  - Distributiile sunt foarte asemanatoare pentru toate categoriile, inclusiv Standard, Low si Moderate.
- Ce suspiciuni/idei putem formula?
  - Folosirea cupoanelor este un comportament comun si nu diferentiaza clar tipurile de clienti. Este posibil ca toti clientii sa fie expusi in mod egal la cupoane sau ca acestea sa fie percepute ca un avantaj usor accesibil.
- Ce preprocesari ar trebui sa aplicam?
  - Standardizare sau normalizare a variabilei cupoane_de_reducere (dupa codificare numerica 0/1).

2.9 **`recenzii <-> tip_client`**



- Ce observam ?
  - Clientii Ultra Vip si Moderate au un numar mare de recenzii, cu distributii concentrate intre 6 si 10.
  - Standard si Vip sunt moderati, cu o dispersie echilibrata intre 2 si 8.
  - Low are cele mai putine recenzii, cu valori frecvente intre 0 si 4.
- Ce suspiciuni/idei putem formula?
  - Clientii Moderate si Ultra Vip sunt cei mai activi in a lasa recenzii, ceea ce poate sugera fie o implicare constanta (Moderate), fie una foarte loiala (Ultra Vip). Clientii Low interactioneaza putin cu magazinul.
- Ce preprocesari ar trebui sa aplicam?
  - Standardizare sau normalizare a variabilei recenzii.

g) Comentarii si interpretări personale

Le-am realizat la fiecare grafic!

7. Antrenarea si evaluarea unui model de baza

-> Am realizat si encodarea si normalizarea/standardizare deoarece eram foarte focusat in laboratoare si nu am realizat ca de fapt acest lucru se face in Partea II.

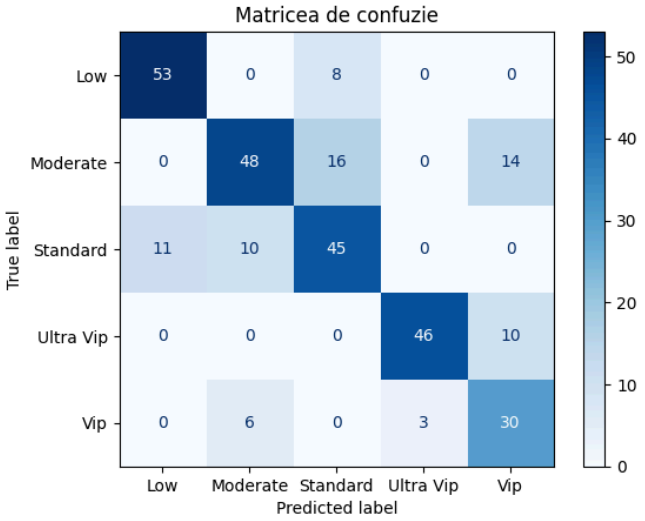
-> Am impartit datele in antrenament si test.

-> Am ales modelul “LogisticRegression” pentru clasificare.

-> Am prezis primele 15 rezultate.

-> Am afisat cele 4 variante de a vedea performanta modelului, toate fiind ~0.74.

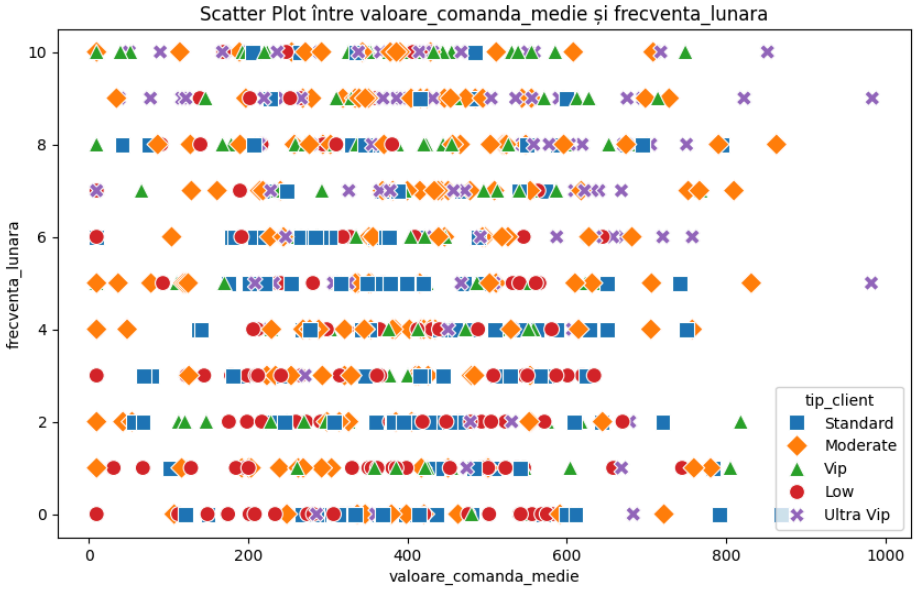
-> Matricea de confuzie:



- Ce observam ?
  - Modelul clasifica foarte bine clasele Low (53 corecte din 61) si Ultra Vip (46 din 56). In schimb, clasele Moderate si Standard se confunda frecvent intre ele (16 instate din Moderate sunt prezise ca Standard, iar 10 din Standard ca Moderate). De asemenea, Moderate este confundat si cu Vip (14 instante), iar Vip este usor confundat cu Moderate si Ultra Vip.
- Ce suspiciuni/idei putem formula?
  - Exista o suprapunere de caracteristici intre clasele Moderate, Standard si Vip, ceea ce sugereaza ca aceste clase sunt greu de diferentiat in functie de trasaturile din setul de date. Este posibil ca unele coloane sa nu ofere destula informatie relevanta pentru a separa aceste categorii. Modelul pare sa distinga clar extremele (Low si Ultra Vip), dar nu si clasele intermediare.
- Ce preprocesari ar trebui sa aplicam?
  - Standardizare a coloanelor numerice, pentru a reduce efectele variatiei scalei intre trasaturi.
  - Normalizare daca exista diferente mari de valori intre coloane, pentru a evita favorizarea unora in model.
  - Reechilibrarea claselor daca unele sunt subreprezentate.
  - Analiza trasaturilor si eliminarea celor irelevante sau adaugarea altora care pot separa mai bine clasele Moderate, Standard si Vip.

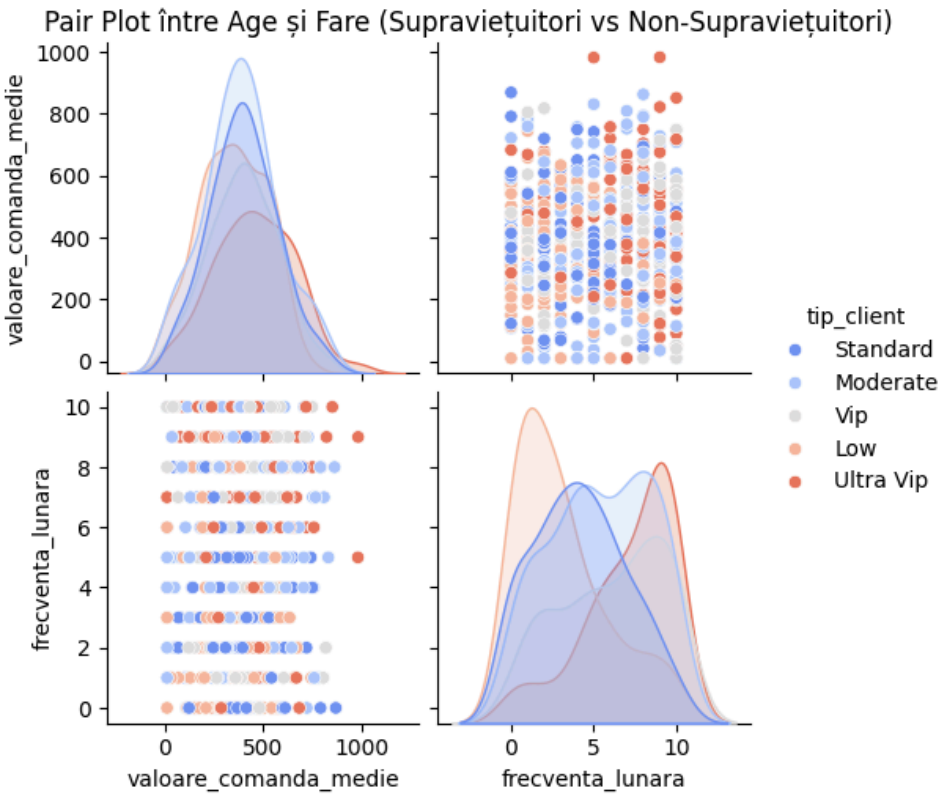
-> Grafice de erori:

1. Scatter plot



- Ce observam ?
  - Clientii din toate categoriile sunt distribuiti pe intreaga plaja a valorii comenzii medii si a frecventei lunare. Totusi, clasele Low si Standard tind sa apara mai frecvent la valori mici ale comenzii medii. In schimb, clasele Moderate si Ultra Vip apar mai des la valori mari. Exista o suprapunere evidenta intre tipurile de clienti, in special in intervalul 200-600 valoare_comanda_medie si frecventa intre 2 si 8.
- Ce suspiciuni/idei putem formula?
  - Este posibil ca trasaturile valoare_comanda_medie si frecventa_lunara sa nu fie suficiente singure pentru a separa tipurile de clienti. Suprapunerea indica faptul ca modelul poate avea dificultati in a face distinctia intre clasele Moderate, Vip si Ultra Vip pe baza acestor trasaturi.
- Ce preprocesari ar trebui sa aplicam?
  - Standardizare a valorii comenzii medii si frecventei lunare pentru a uniformiza scara.
  - Normalizare pentru a reduce impactul valorilor extreme.
  - Adaugarea de noi trasaturi derivate (de exemplu, total_comanda_lunara = valoare_comanda_medie * frecventa_lunara) care ar putea oferi o separare mai clara intre clase.
  - Analizarea corelatiilor pentru a vedea daca aceste trasaturi chiar ajuta la predictie sau trebuie combinate cu altele.

2. Pair plot



- Ce observam ?
  - Distributiile pentru valoare_comanda_medie sunt in general asemanatoare pentru toate clasele, cu un varf in jurul valorii 200-300. Totusi, clasa Ultra Vip pare sa aiba o distributie usor mai larga si mai plata. La frecventa_lunara, clasele Low si Moderate au varfuri clare, dar sunt suprapuse semnificativ cu restul. Distributiile claselor sunt greu de separat vizual.
- Ce suspiciuni/idei putem formula?
  - Trasaturile valoare_comanda_medie si frecventa_lunara nu reusesc sa separe clar clasele de clienti. Suprapunerea indica faptul ca aceste atribute nu contin informatii suficiente pentru clasificare. Posibil ca lipsesc trasaturi suplimentare (de exemplu: istoric achizitii, interactiuni cu aplicatia etc.) care sa ajute la distinctie.
- Ce preprocesari ar trebui sa aplicam?
  - Standardizare pentru a face trasaturile comparabile.
  - Normalizare pentru a reduce efectul valorilor extreme.
  - Creare de trasaturi combinate (ex: scor_client = valoare_comanda_medie * frecventa_lunara).

FINAL!