Proiect PCLP3

Duță Ștefan-Horia, 311CA

Descriere:

1. Tipul problemei:

   Setul de date este destinat unei probleme de regresie, care consta in predictia veniturilor unei companii.

2. Structura setului de date:

   Setul de date va avea 2000 de exemple si este impartit in doua subseturi:

   - Subsetul de antrenare: are 1400 de exemple
   - Subsetul de testare: are 600 de exemple

3. Caracteristici:

   Fiecare instanta are 9 coloane relevante, dintre care “venit” este cea tinta. Pentru acestea am folosit, 3 tipuri diferite de date: numere intregi(numar angajati, an infiintare, numar produse, numar clienti, numar parteneri), numere reale(capital social, venit) si valori categoriale(domeniu de activitate, tara)

4. Salvarea dataseturilor:

   Subsetul de antrenare va fi salvat in “train.csv”, iar cel de testare in “test.csv”.

5. Explicarea modului de constructie a setului de date:

   Problema noastra consta in predictia veniturilor unei companii. Setul de date a fost generat aleator, folosind distributia normala si uniforma, cu urmatoarele restrictii: media numarului de angajati este de 100, capitalul social este intre 50000 si 500000 de euro, anul de infiintare este intre 1970 si 2024, numarul de produse este intre 1 si 100, numarul de clienti este intre 50 si 1000, iar numarul de parteneri intre 1 si 20. Tara si domeniul de activitate au fost alese aleator dintr-un set fix de optiuni, acestea fiind Romania, Germania, Franta, Italia, Spania, respectiv, IT, Retail, Agricultura, Educatie, Financiar. Venitul a fost calculat folosind o formula la care contribuie toate caracteristicile unei instante si anume:

   (numar_angajati * 3000 + capital_social * 0.5 + vechime * 1000 + numar_clienti * 500 + numar_parteneri * 1000) * coef_domeniu * coef_tara.

   Apoi, alegem un numar intre 50 si 100 de valori din fiecare coloana pentru a le sterge.

6. Analiza exploratorie a datelor (EDA complex):

   Analizare valori lipsa:

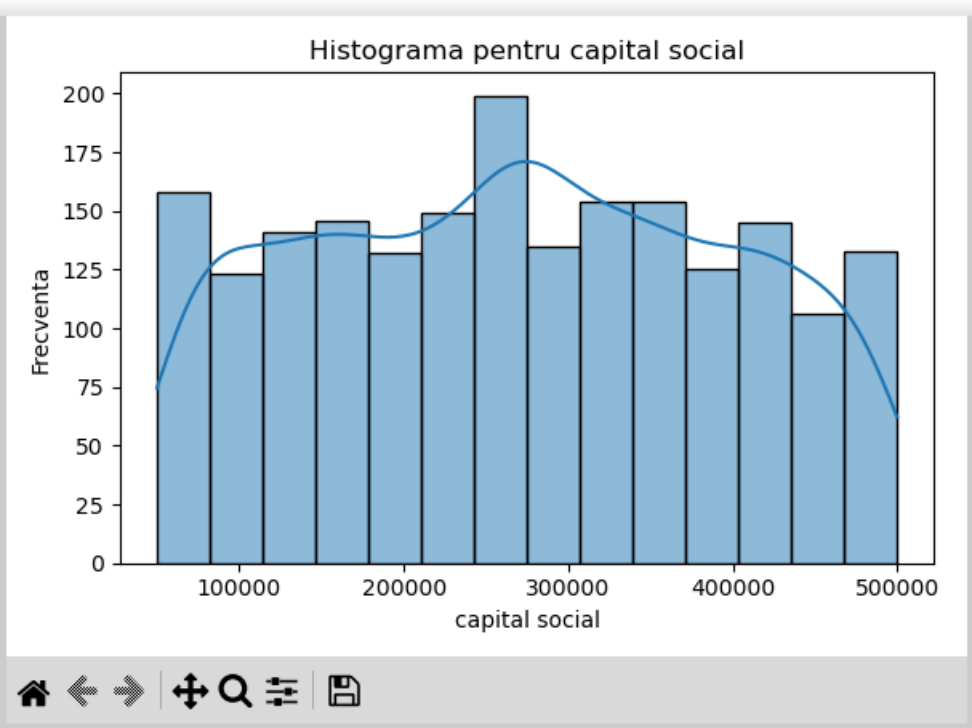
numar angajati	54
capital social	83
an infiintare	71
numar produse	70
domeniu de activitate	94
tara	86
clienti	58
parteneri	67
venit	0

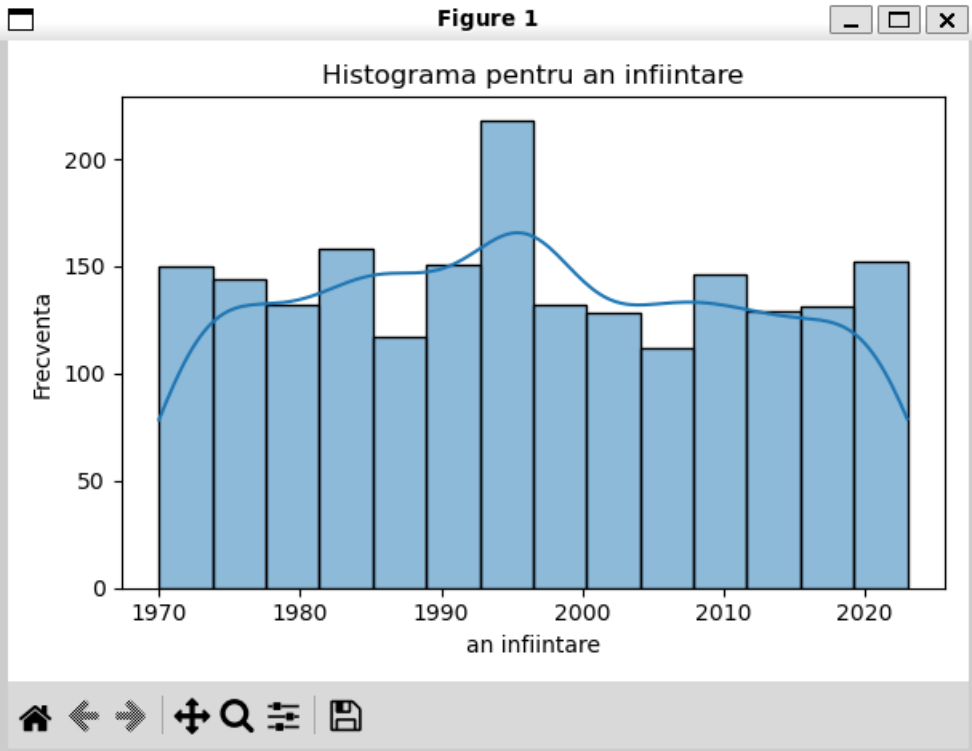
   In cazul coloanelor cu valori numerice, vom inlocui spatiile libere cu media aritmetica de pe acea coloana, iar in cazul celor cu valori categoriale, vom inlocui cu cea mai frecventa eticheta.

Statistici descriptive:

	numar angajati	capital social	an infiintare	numar produse	clienti	parteneri	venit
count	2000.000000	2000.000000	2000.000000	2000.000000	2000.000000	2000.000000	2.000000e+03
mean	100.405961	270572.452888	1996.191291	50.118653	533.848095	10.188308	1.020487e+06
std	19.442187	126352.652747	15.374407	28.087174	267.425996	5.337319	3.590860e+05
min	35.000000	50005.235640	1970.000000	1.000000	50.000000	1.000000	2.714512e+05
25%	87.000000	163007.283415	1984.000000	27.000000	305.000000	6.000000	7.477798e+05
50%	100.405961	270572.452888	1996.191291	50.118653	533.848095	10.188308	9.756660e+05
75%	113.000000	373262.840762	2009.000000	73.000000	758.000000	15.000000	1.234982e+06
max	177.000000	499800.966463	2023.000000	99.000000	999.000000	19.000000	2.219132e+06

Analiza distribuitiei variabilelor:





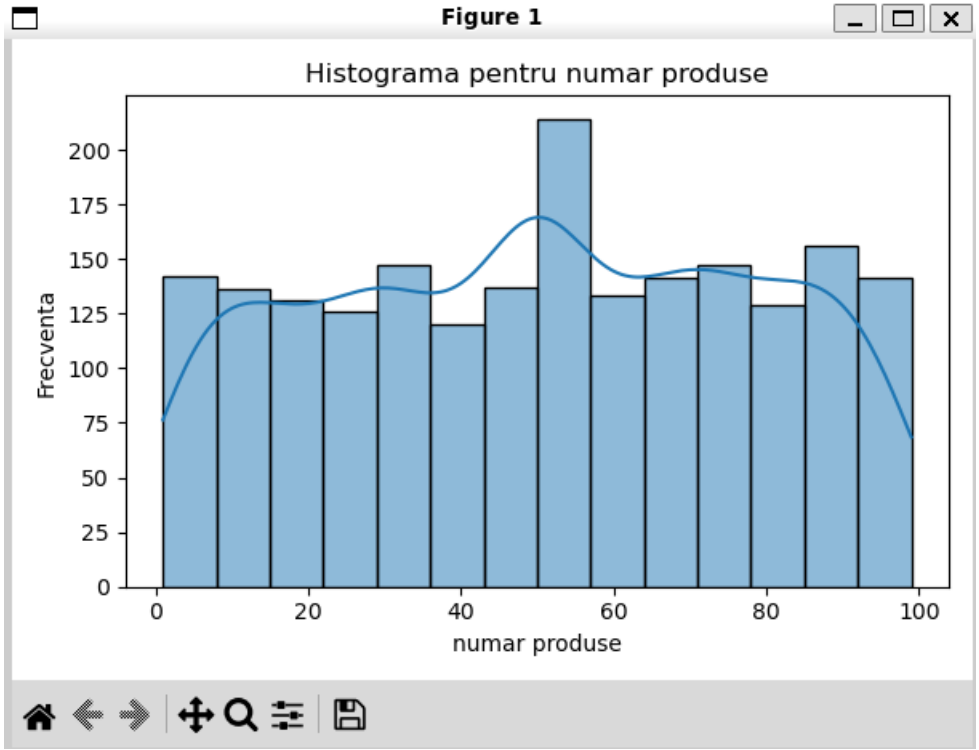


Figure 1
Histograma pentru numar produse
Frecventa
numar produse
0
20
40
60
80
100
0
25
50
75
100
125
150
175
200

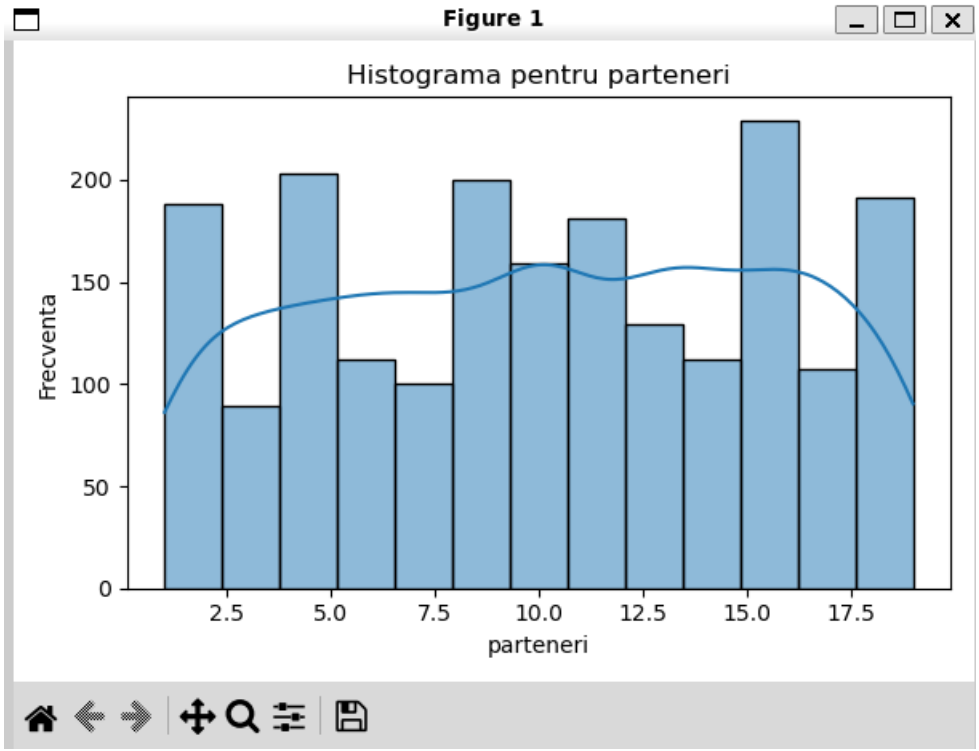


Figure 1
Histograma pentru parteneri
Frecventa
parteneri
2.5
5.0
7.5
10.0
12.5
15.0
17.5
0
50
100
150
200

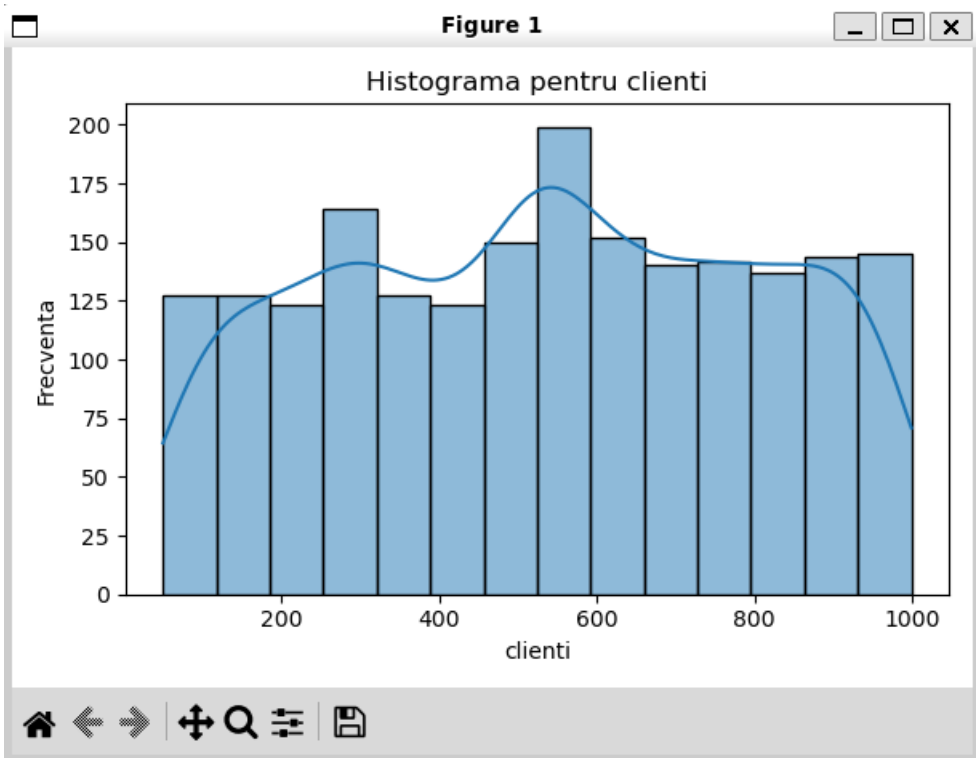


Figure 1
Histograma pentru clienti
Frecventa
clienti
0
25
50
75
100
125
150
175
200
200
400
600
800
1000

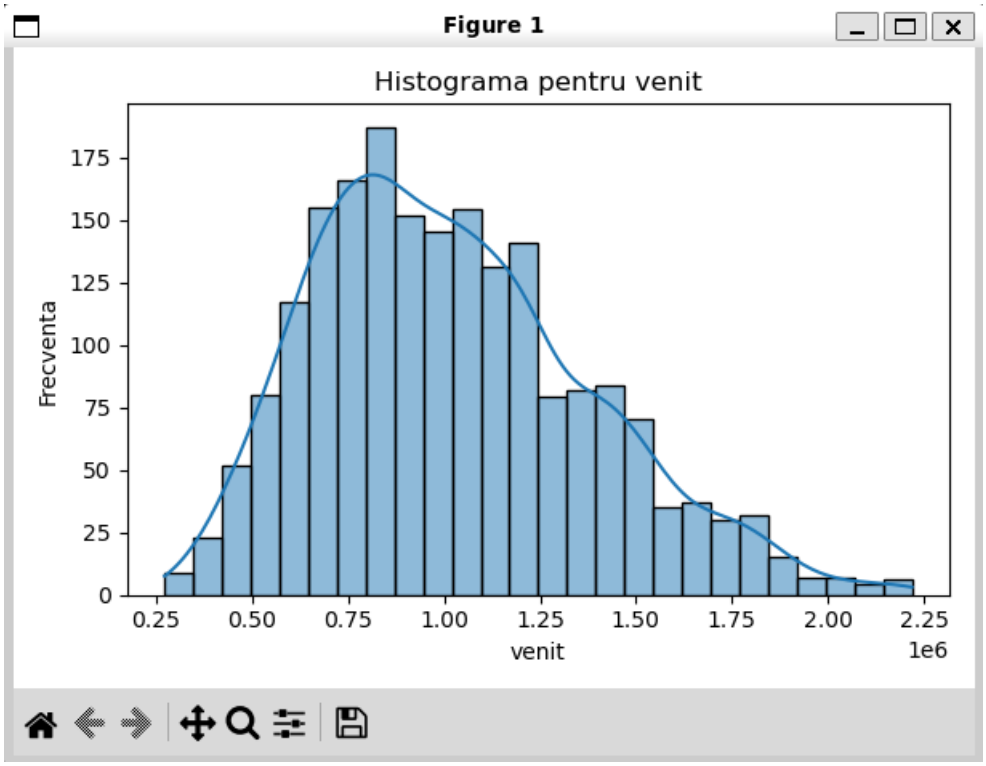


Figure 1
Histograma pentru venit
Frecventa
venit
1e6
0
25
50
75
100
125
150
175
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00
2.25

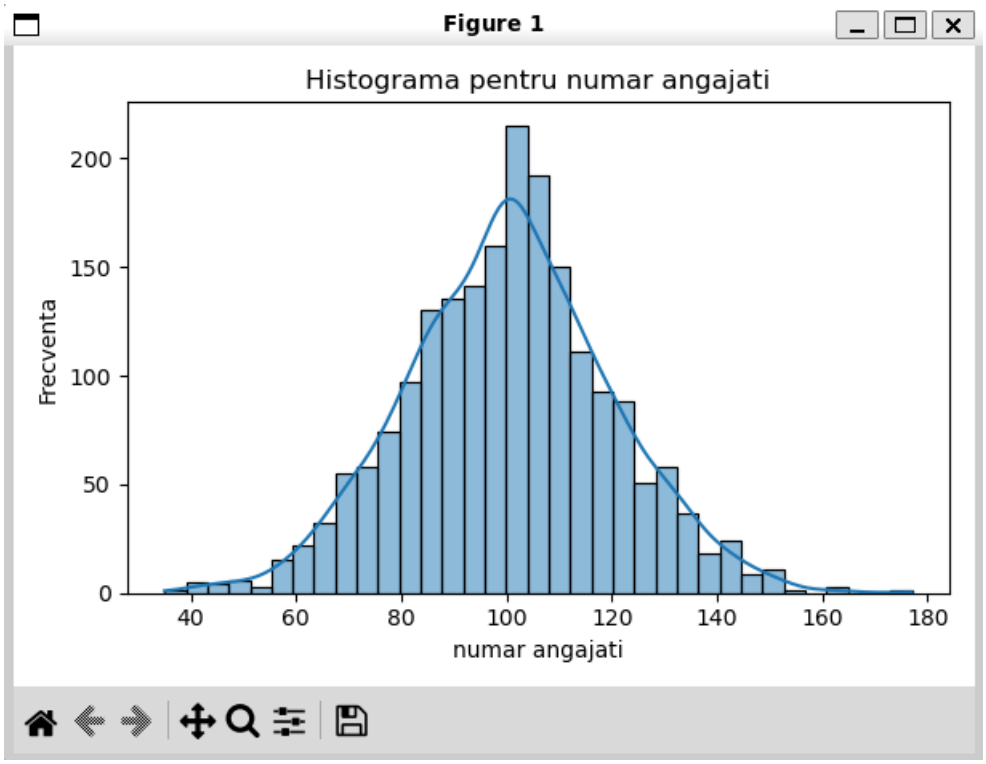
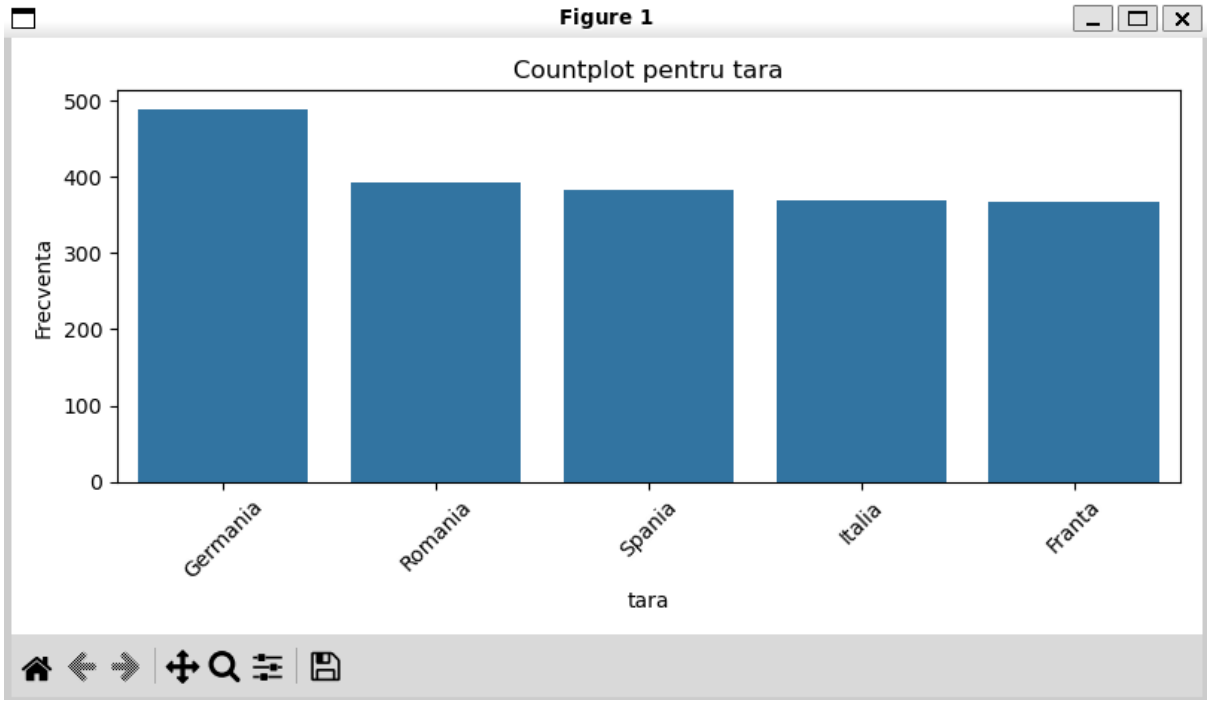
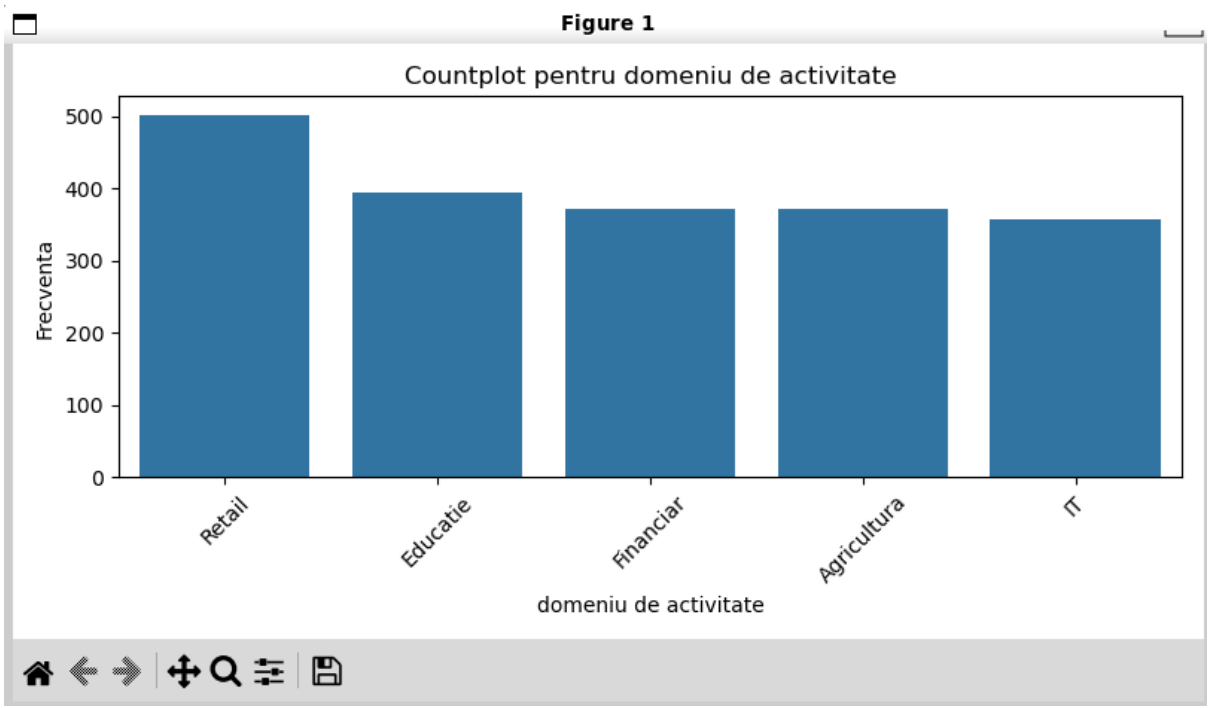


Figure 1
Histograma pentru numar angajati
Frecventa
numar angajati
0
50
100
150
200
40
60
80
100
120
140
160
180





Analiza corelatiilor:

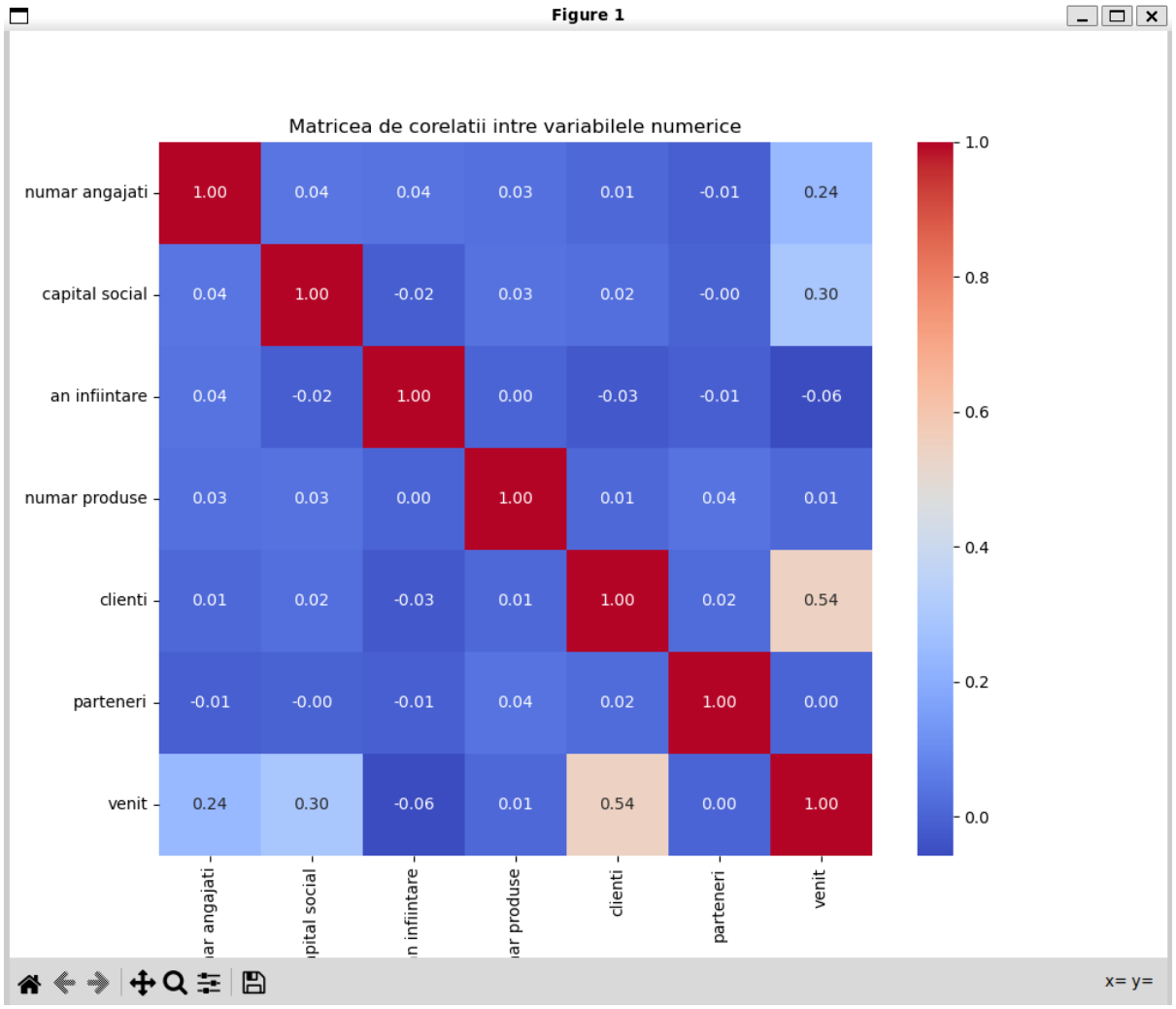
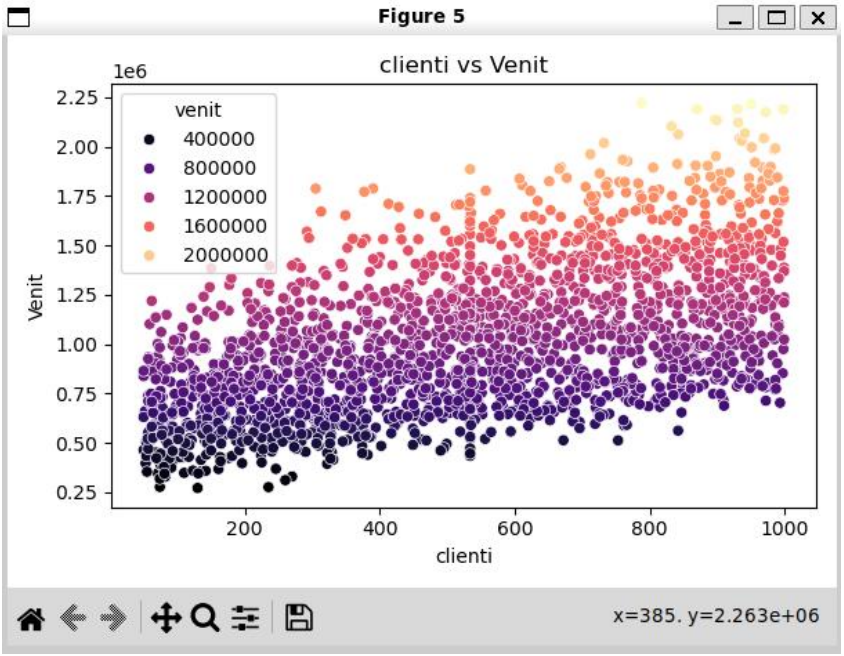
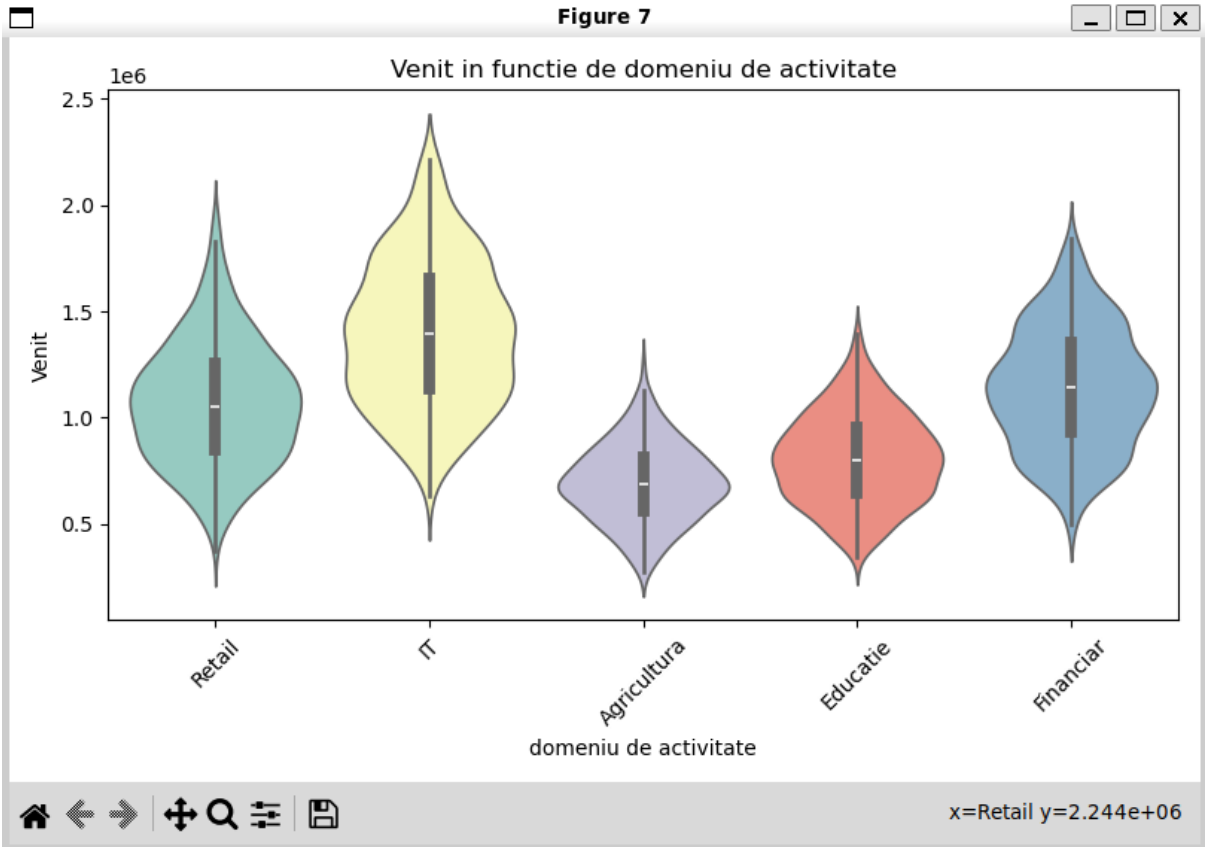


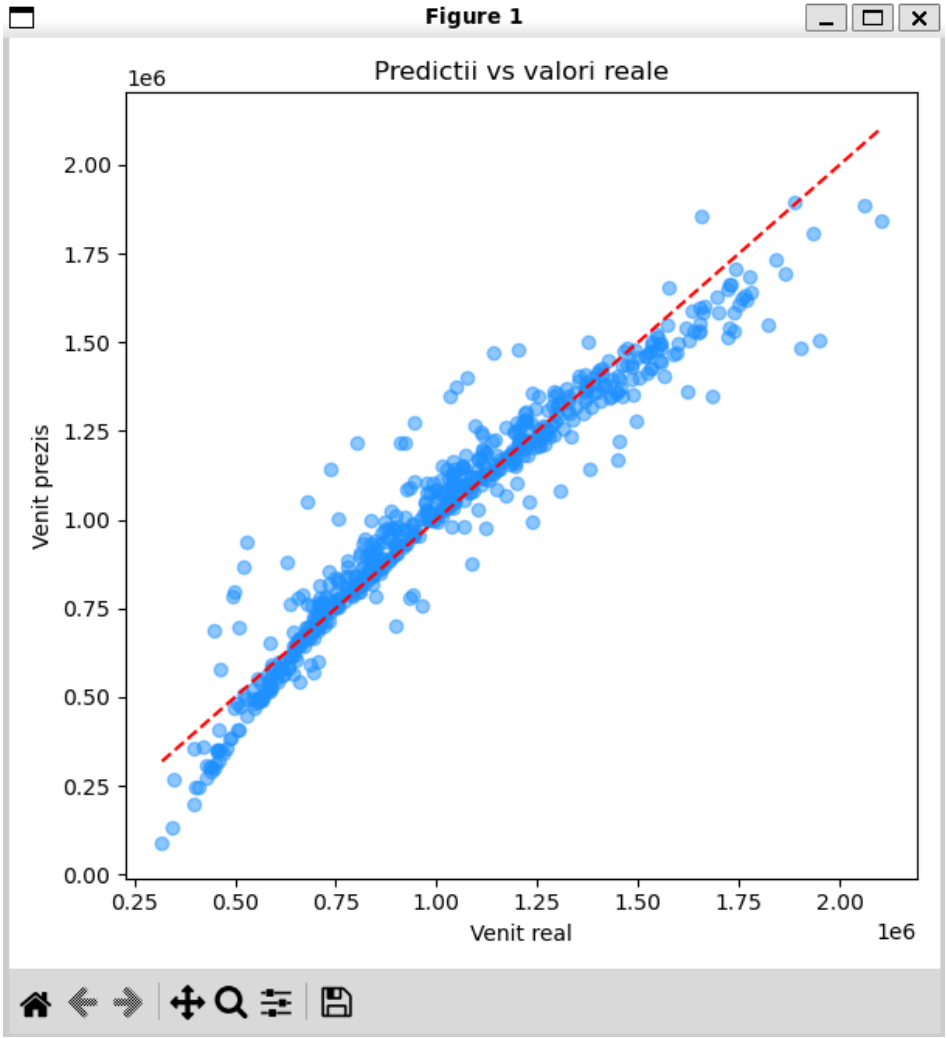
Figure 1
Matricea de corelatii intre variabilele numerice
numar angajati
capital social
an infiintare
numar produse
clienti
parteneri
venit
1.00 0.04 0.04 0.03 0.01 -0.01 0.24
0.04 1.00 -0.02 0.03 0.02 -0.00 0.30
0.04 -0.02 1.00 0.00 -0.03 -0.01 -0.06
0.03 0.03 0.00 1.00 0.01 0.04 0.01
0.01 0.02 -0.03 0.01 1.00 0.02 0.54
-0.01 -0.00 -0.01 0.04 0.02 1.00 0.00
0.24 0.30 -0.06 0.01 0.54 0.00 1.00
1.0
0.8
0.6
0.4
0.2
0.0
x= y=

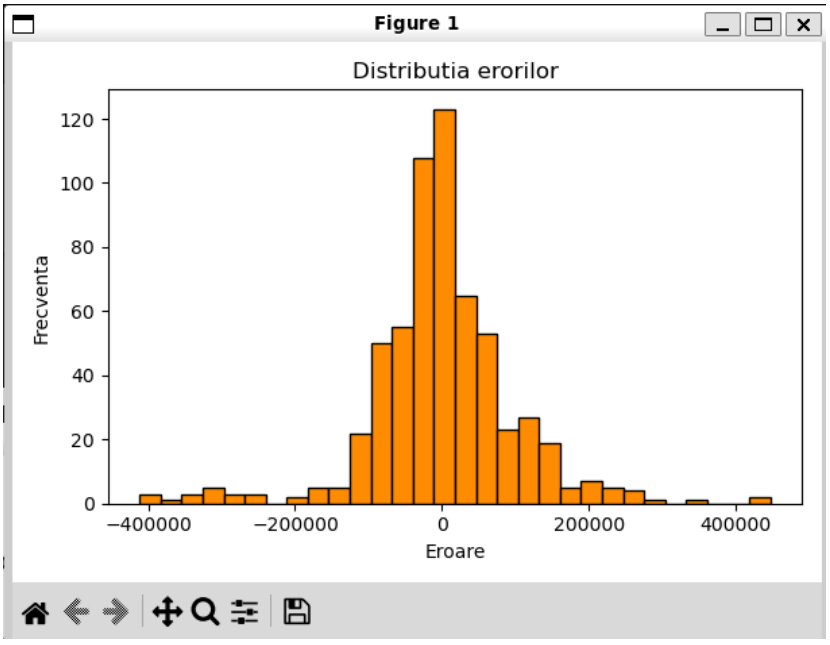
Analiza relatiilor cu variabila tinta:





7. Antrenarea si evaluarea unui model de baza:






```
MAE: 65922.13
RMSE: 98376.38
R^2: 0.9205
```