Obiectivele lucrării

Obiectivul principal al acestei lucrări este de a utiliza regresia liniară pentru a prezice sumele facturate pe baza datelor istorice de vânzări. În acest sens, vom urmări următoarele obiective specifice:

- Pregătirea şi explorarea setului de date de vânzări.

- Implementarea unui model de regresie liniară pentru a prezice sumele facturate.

- Evaluarea performanţei modelului şi interpretarea rezultatelor obţinute.

Prin atingerea acestor obiective, ne propunem să evidenţiem utilitatea şi eficacitatea regresiei liniare în anticiparea cererii şi în luarea deciziilor într-un mediu comercial.

Metodologie

În această secţiune, vom oferi o descriere detaliată a metodelor şi tehnicilor utilizate în lucrare pentru pregătirea datelor, implementarea modelului de regresie liniară şi evaluarea performanţei acestuia.

- **Pregătirea datelor**

  Pregătirea datelor este un pas esenţial în procesul de dezvoltare a modelului. În acest scop, am efectuat următoarele etape:

  1. Importul datelor: Am încărcat setul de date de vânzări într-un cadru de date utilizând biblioteca pandas în Python.

  2. Explorarea datelor: Am analizat şi am explorat setul de date pentru a înţelege structura acestuia, tipurile de date ale fiecărei variabile şi pentru a identifica eventualele probleme sau lipsuri.

  3. Pregătirea datelor: Am efectuat operaţiuni de curăţare a datelor, inclusiv tratarea valorilor lipsă, eliminarea sau înlocuirea datelor anormale şi transformarea datelor non-numerice într-o formă numerică utilizabilă pentru model.

- **Modelul de regresie liniară**

  Modelul de regresie liniară este utilizat pentru a estima relaţia liniară între variabila dependentă (suma facturată) şi una sau mai multe variabile independente (baza impozitare, valoare TVA). În implementarea noastră, am urmat următoarele etape:

  1. Definirea variabilelor dependente şi independente: Am selectat variabilele relevante din setul de date pentru a fi utilizate în model.

  2. Împărţirea datelor: Am împărţit datele într-un set de date de antrenament şi un set de date de testare pentru a evalua performanţa modelului pe date nevăzute.

  3. Antrenarea modelului: Am utilizat setul de date de antrenament pentru a antrena modelul de regresie liniară.

4. Testarea modelului: Am testat performanţa modelului folosind setul de date de testare şi am evaluat metrici cum ar fi Mean Squared Error (MSE) şi R^2 Score.

- **Evaluarea modelului**

Pentru a evalua performanţa modelului, am folosit următoarele metrici:

- Mean Squared Error (MSE): Este o măsură a dispersiei dintre valorile reale şi cele prezise ale variabilei dependente. Cu cât este mai mic MSE, cu atât modelul este considerat mai precis.

- R^2 Score: Reprezintă coeficientul de determinare şi oferă o măsură a calităţii ajustării modelului la datele de antrenament. Valoarea R^2 apropiată de 1 indică o bună potrivire a modelului la date.

- **Predicţia**

Pentru a realiza predicţii cu privire la sumele facturate în viitor, am utilizat modelul de regresie liniară antrenat pe datele existente. Procesul de predicţie a urmat următoarele etape:

Selectarea caracteristicilor: Am selectat caracteristicile relevante care să fie utilizate pentru a face predicţiile. În cazul nostru, aceste caracteristici au inclus baza impozitare şi valoarea TVA.

Pregătirea datelor de predicţie: Am pregătit datele de predicţie, asigurându-ne că acestea au aceeaşi structură şi formă ca şi datele folosite pentru antrenamentul modelului.

Realizarea predicţiilor: Am aplicat modelul antrenat pe datele de predicţie pentru a obţine estimări cu privire la sumele facturate.

Evaluarea predicţiilor: Am comparat predicţiile obţinute cu datele reale existente pentru a evalua acurateţea şi performanţa modelului în a prezice sumele facturate.

Rezultate

Setul de date are o structură tabulară şi este organizat într-un format tip tabel cu rânduri şi coloane. Fiecare rând reprezintă o înregistrare individuală sau o observaţie, în timp ce fiecare coloană reprezintă o variabilă sau un atribut specific.

Structura setului de date include următoarele caracteristici:

- Dimensiuni: 6461 de înregistrări şi 6 coloane.

- Coloane: Fiecare coloană reprezintă o variabilă specifică sau un atribut al datelor. În cazul acestui set de date, coloanele includ:

1. Nr. factura: Numărul facturii
2. Data factura: Data la care a fost emisă factura.
3. Client: Numele sau identificatorul clientului asociat facturii.

4. Suma factura: Valoarea totală a facturii.

5. Baza impozitare: Valoarea de bază a facturii, fără TVA.

6. Valoare TVA: Valoarea taxei pe valoarea adăugată (TVA) asociată facturii.

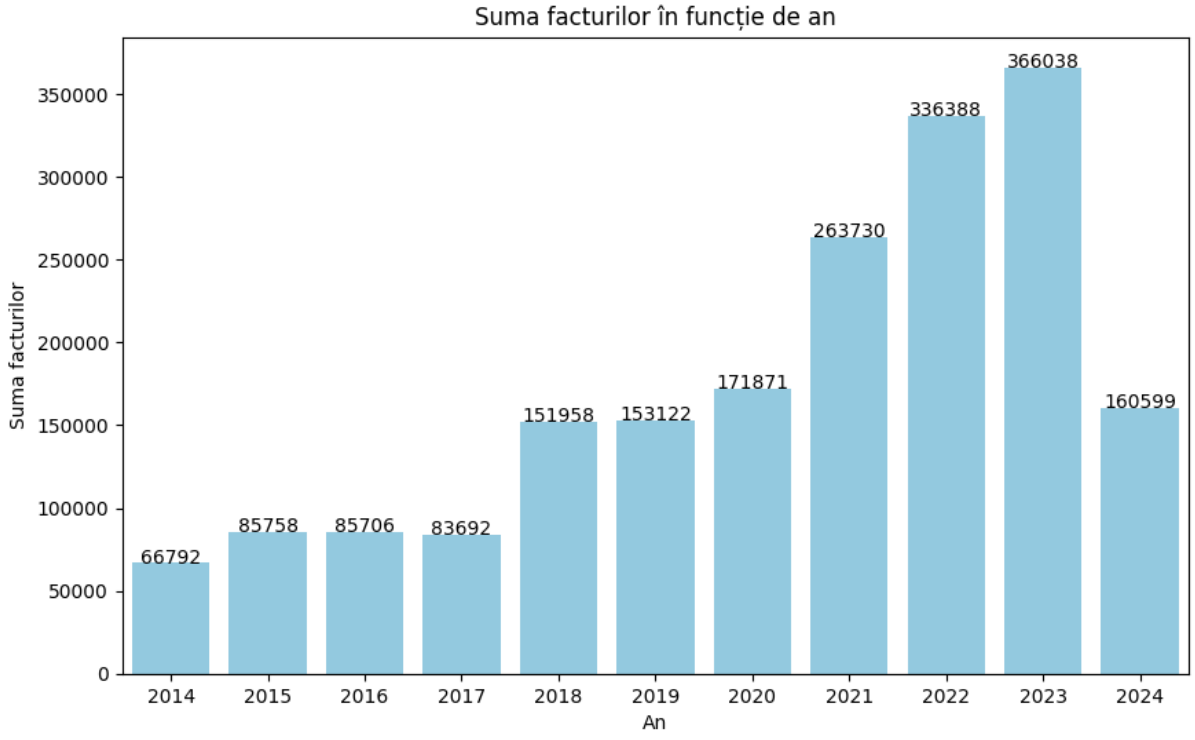
- Tipuri de date: Fiecare coloană poate avea un tip de date specific, cum ar fi întregi, şiruri de caractere sau valori reale. De exemplu, Nr. factura poate fi de tip întreg, Data factura şi Client pot fi şiruri de caractere, iar Suma factura, Baza impozitare şi Valoare TVA pot fi valori reale.

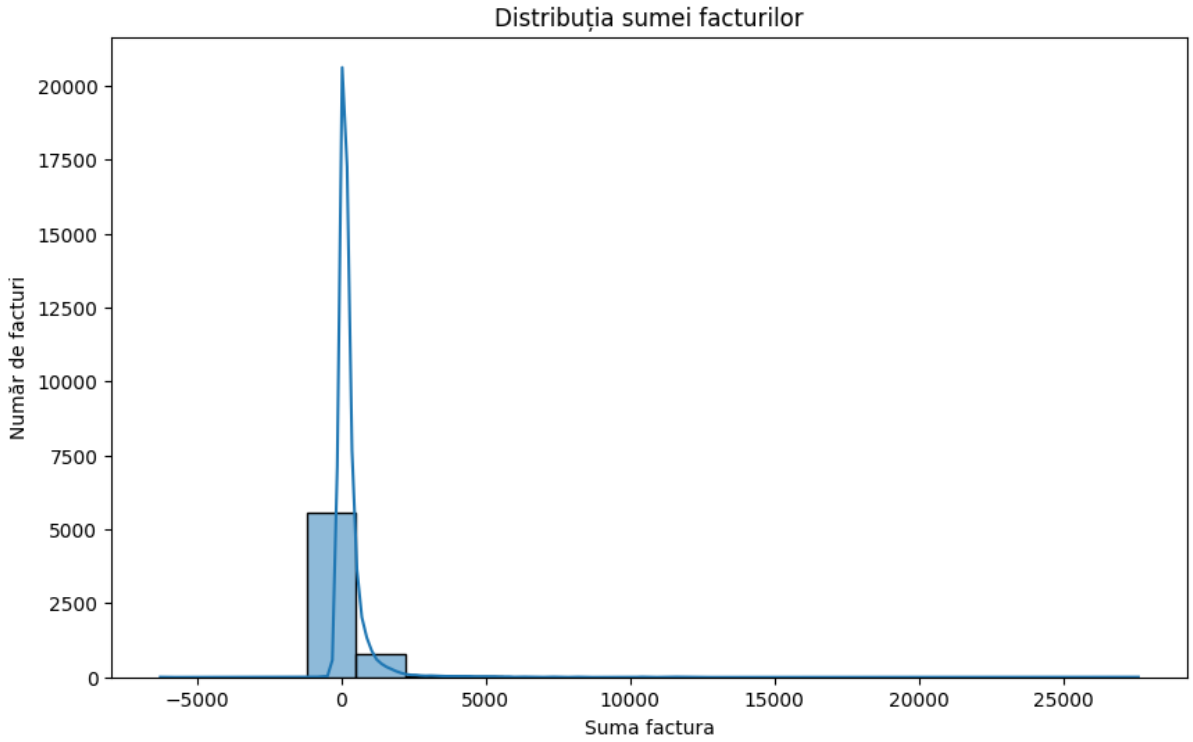
Vizualizare primele 5 rânduri din set:

	Nr. factura	Data factura	Client
0	53962	02.05.2014 00:00:00	MEGA AUTO S.R.L.
1	53963	02.05.2014 00:00:00	GIP EST SRL
2	53964	05.05.2014 00:00:00	COLEGIUL NATIONAL DE ARTA OCTAV BANCILA
3	53971	05.05.2014 00:00:00	EPTISA ROMANIA SRL
4	53972	08.05.2014 00:00:00	PALATUL COPIILOR - IASI

	Suma factura	Baza impozitare	Valoare TVA
0	30.24	24.39	5.85
1	407.51	328.64	78.87
2	56.00	45.16	10.84
3	270.44	218.10	52.34
4	180.00	145.16	34.84

Vizualizare grafice:



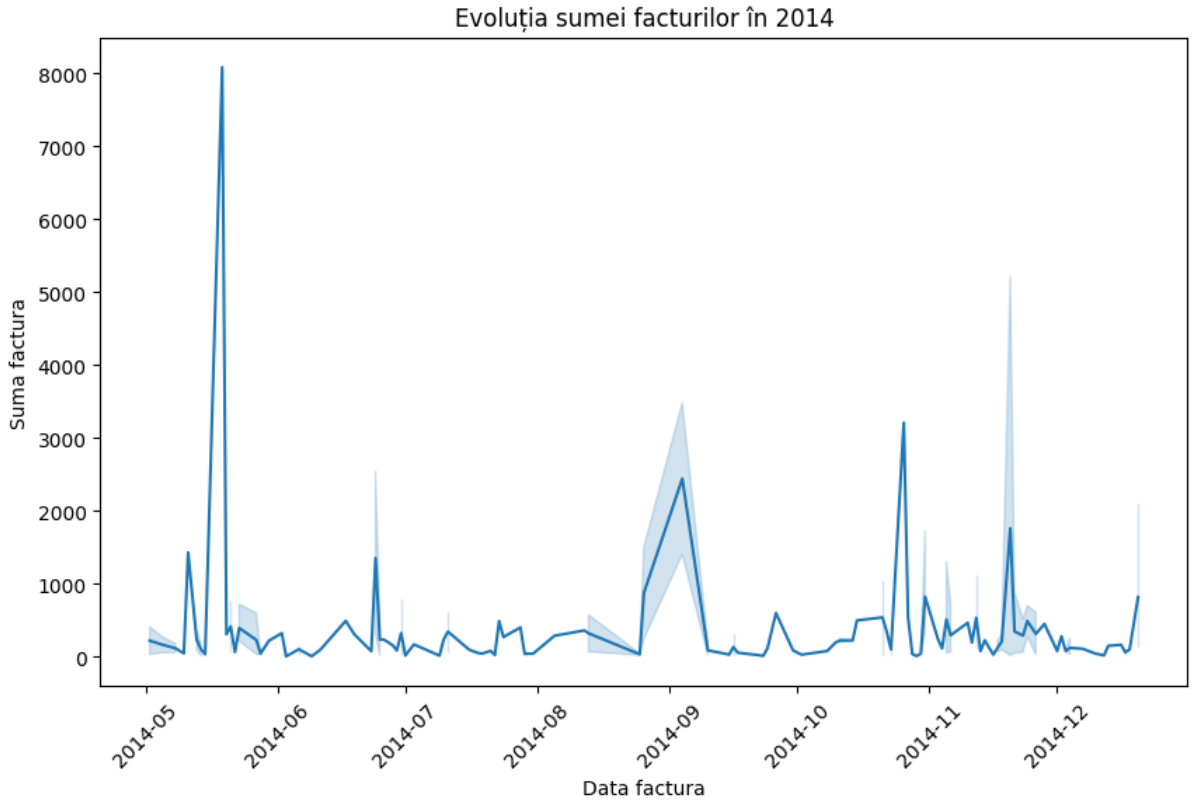


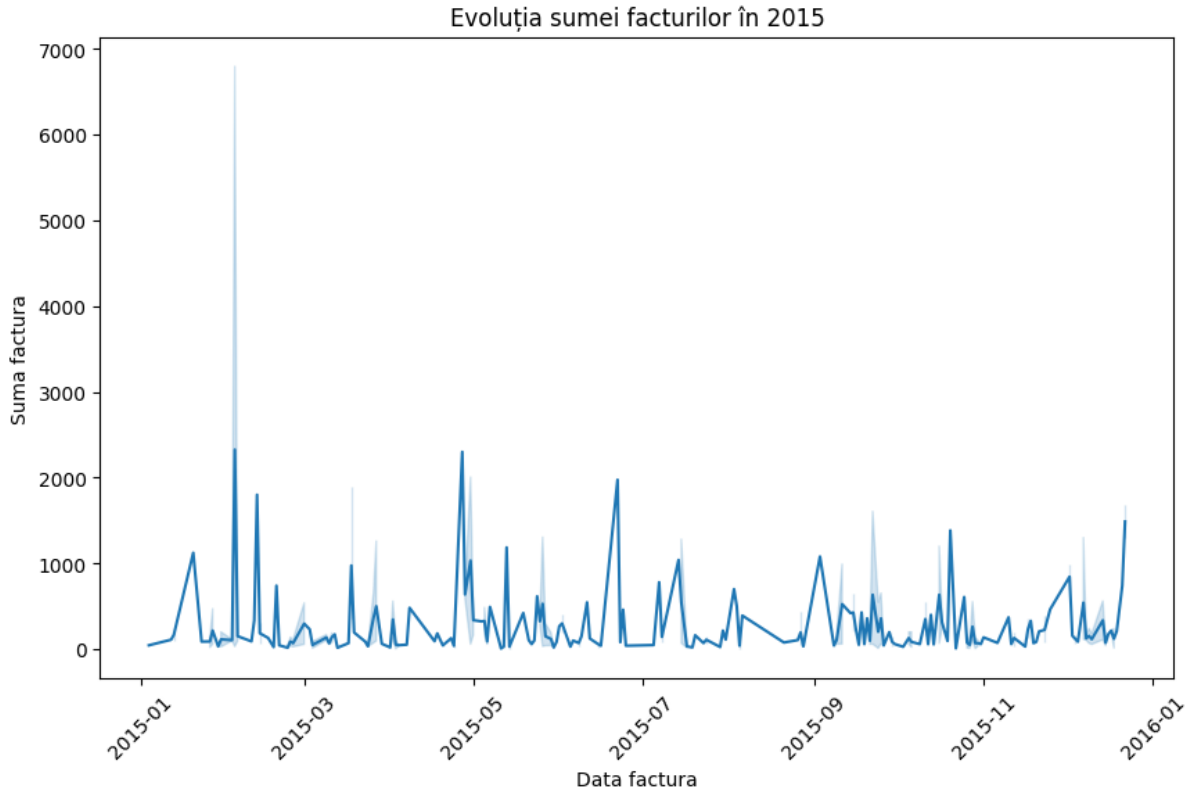
Graficul prezentat arată distribuţia sumei facturilor, având pe axa orizontală sumele facturilor şi pe axa verticală numărul de facturi.

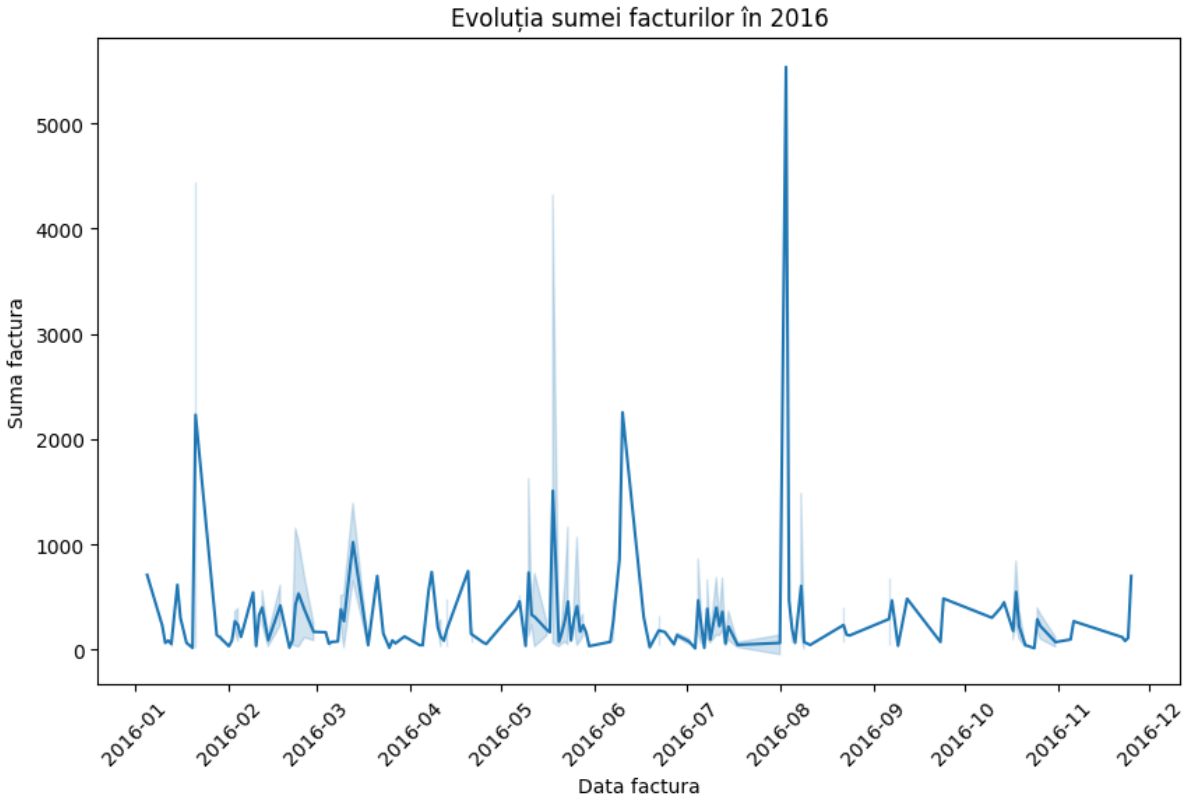
Observăm că majoritatea facturilor se concentrează în jurul valorii zero, unde vedem un vârf foarte pronunţat, indicând o frecvenţă mare a facturilor cu sume foarte mici sau zero. Există de asemenea o mică frecvenţă pentru facturi cu valori negative, posibil reprezentând rambursări sau corecţii.

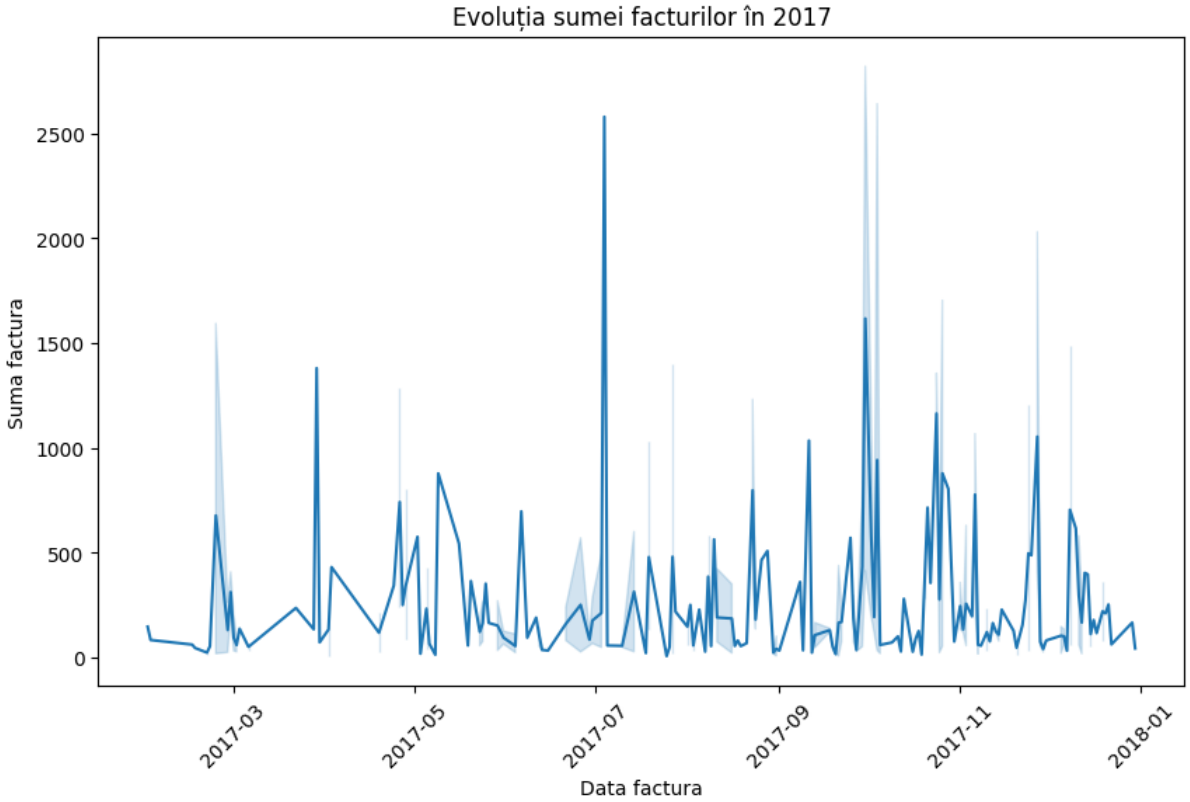
După acest vârf pronunţat, numărul de facturi scade dramatic odată cu creşterea sumei facturii, sugerând că facturile cu sume mari sunt mult mai rare.

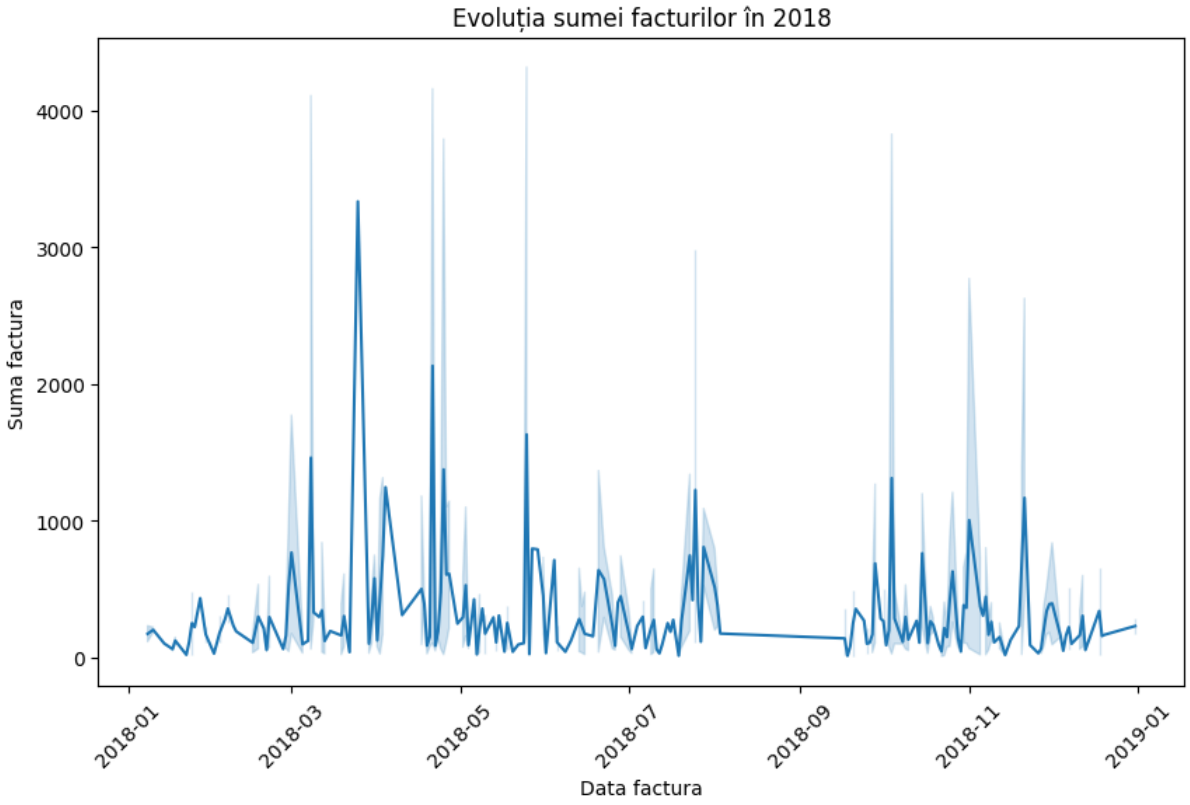
Evolutia sumei facturilor pe ani:

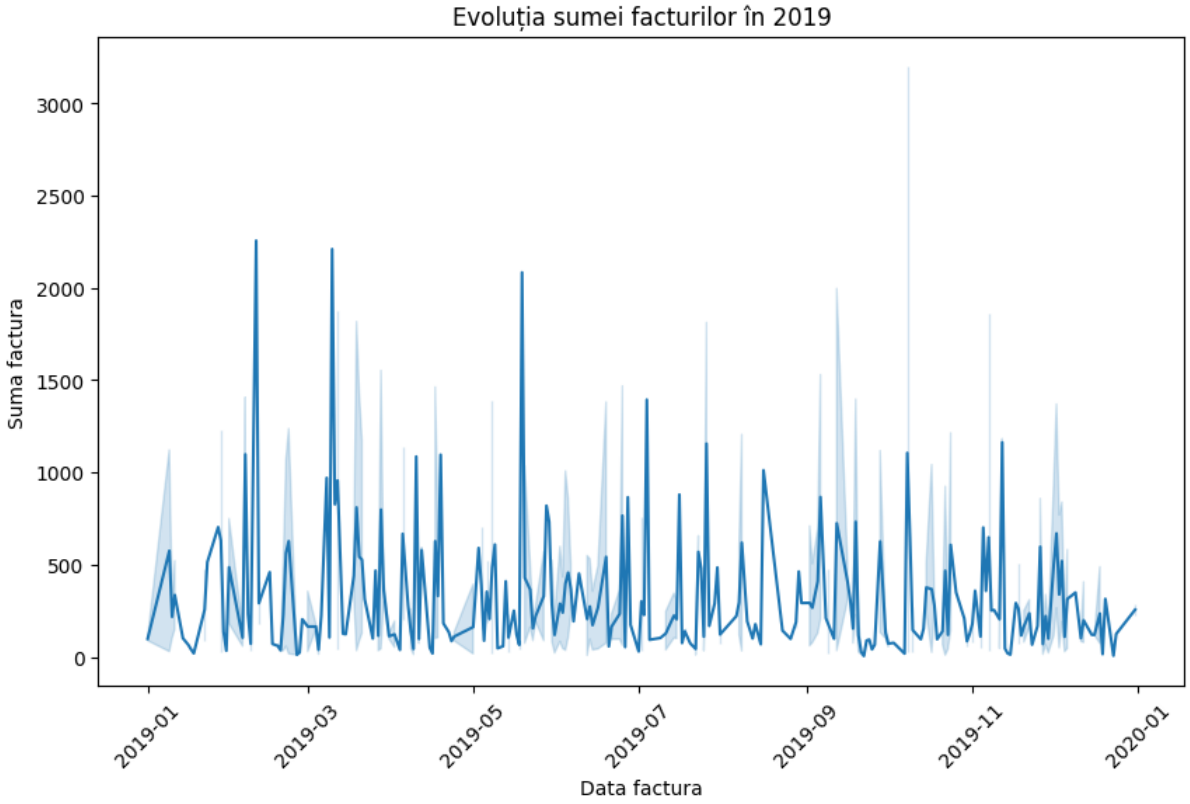


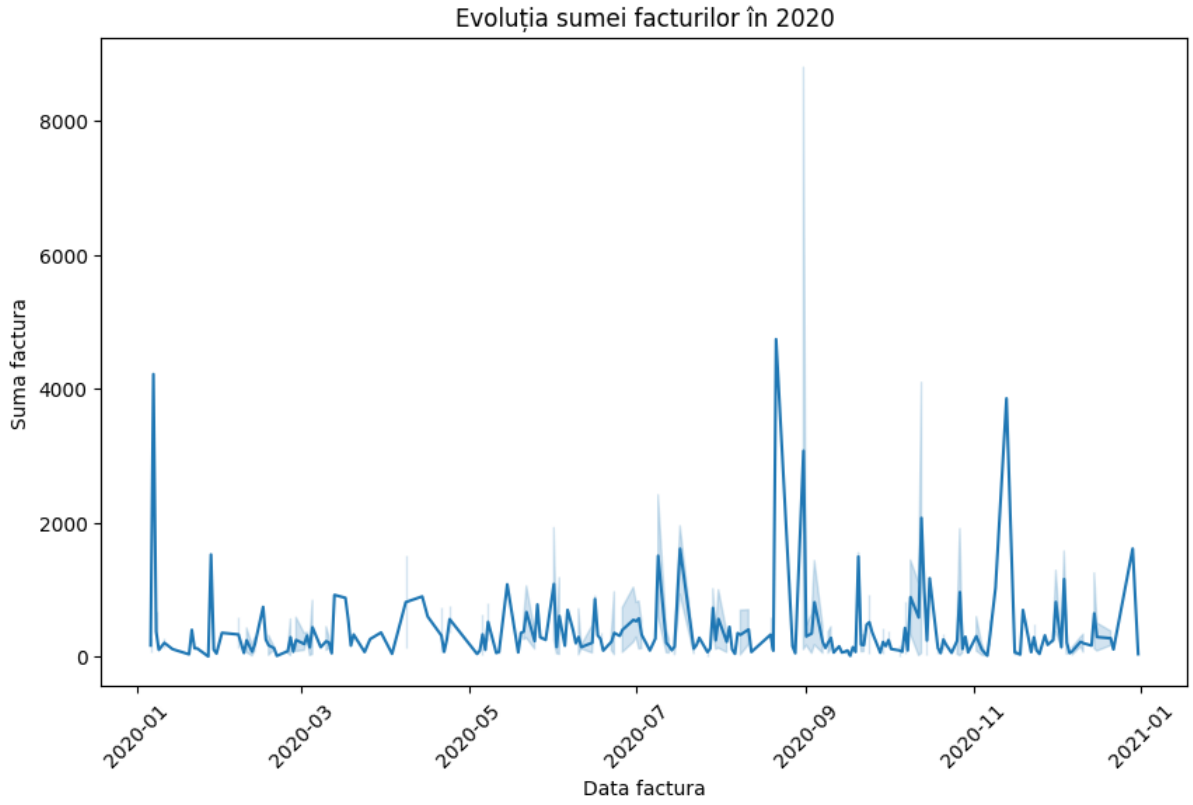


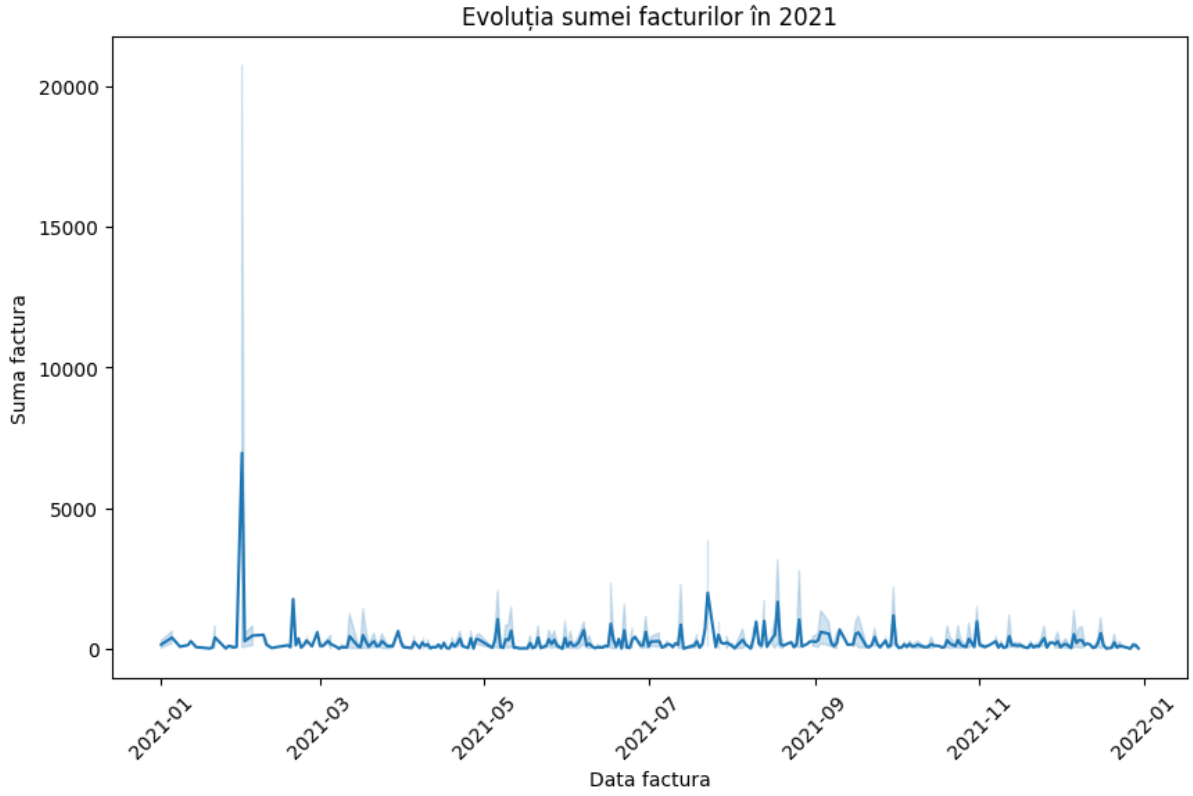

Evoluția sumei facturilor în 2016
Suma factura
0
1000
2000
3000
4000
5000
2016-01
2016-02
2016-03
2016-04
2016-05
2016-06
2016-07
2016-08
2016-09
2016-10
2016-11
2016-12
Data factura

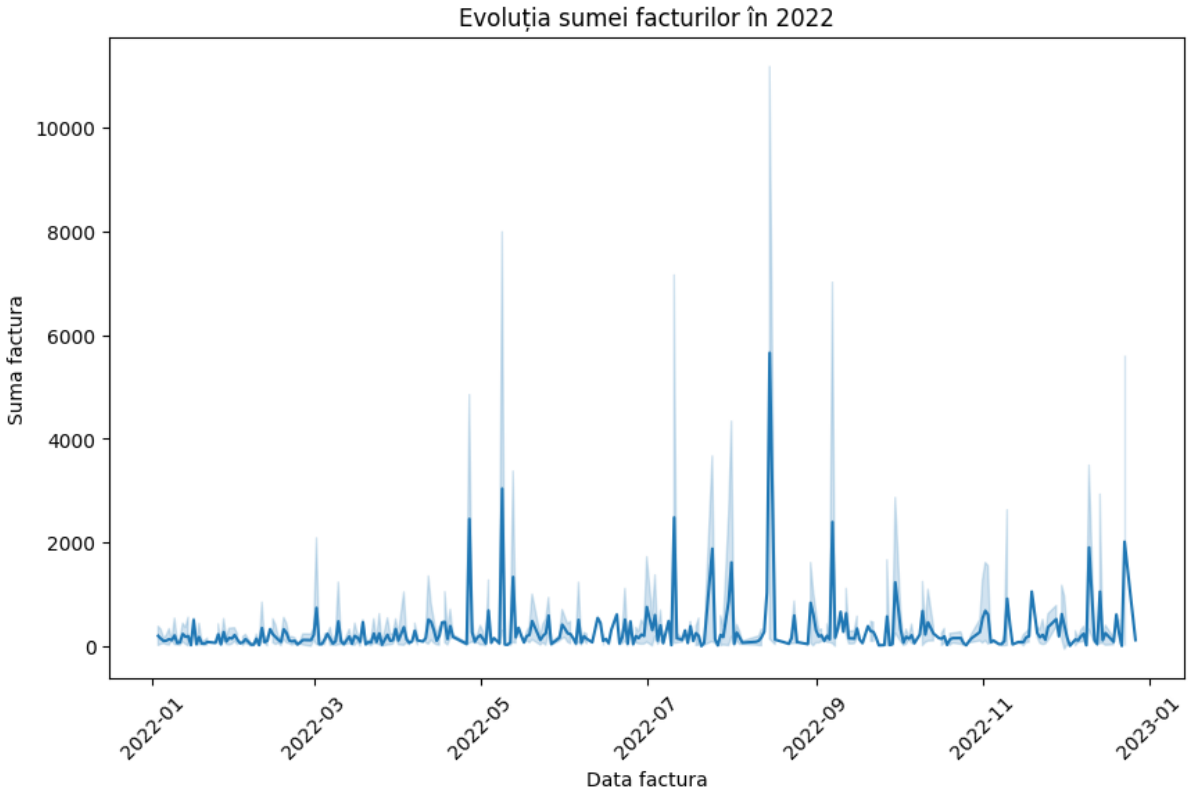

Evoluția sumei facturilor în 2017
Suma factura
0
500
1000
1500
2000
2500
2017-03
2017-05
2017-07
2017-09
2017-11
2018-01
Data factura

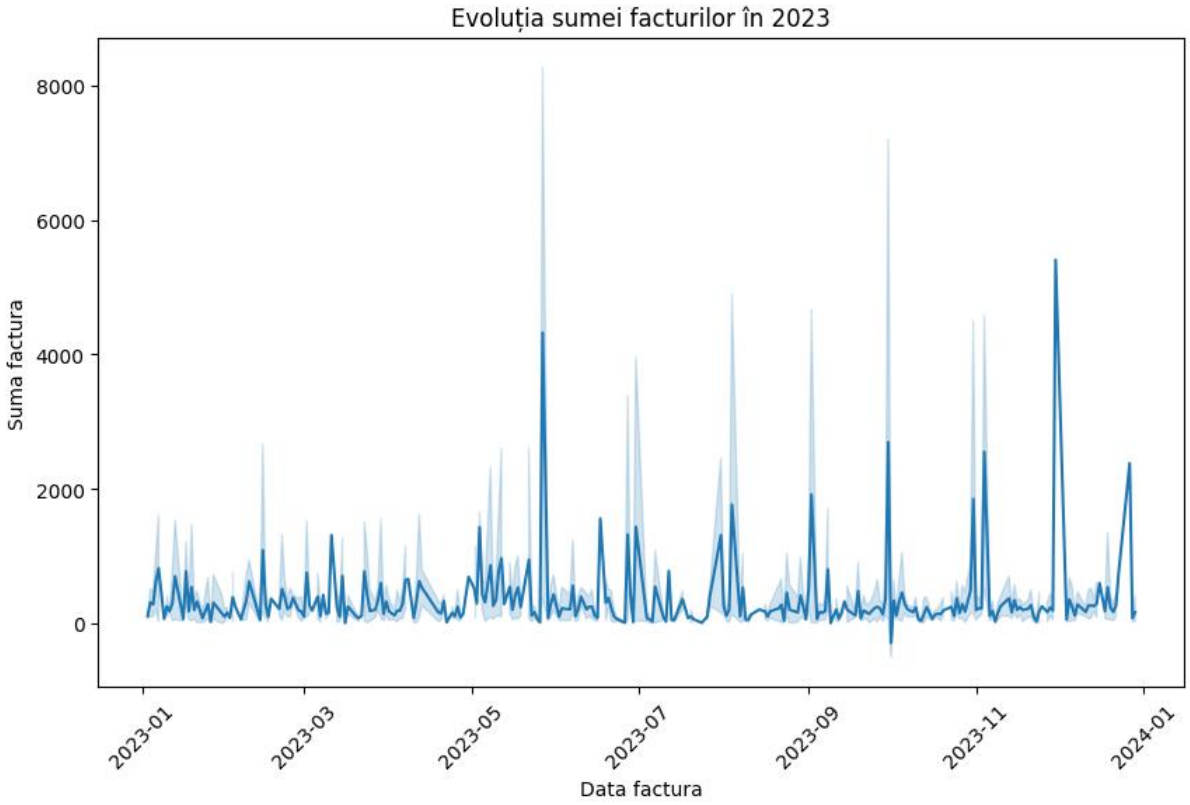

Evoluția sumei facturilor în 2018
Suma factura
0
1000
2000
3000
4000
2018-01
2018-03
2018-05
2018-07
2018-09
2018-11
2019-01
Data factura

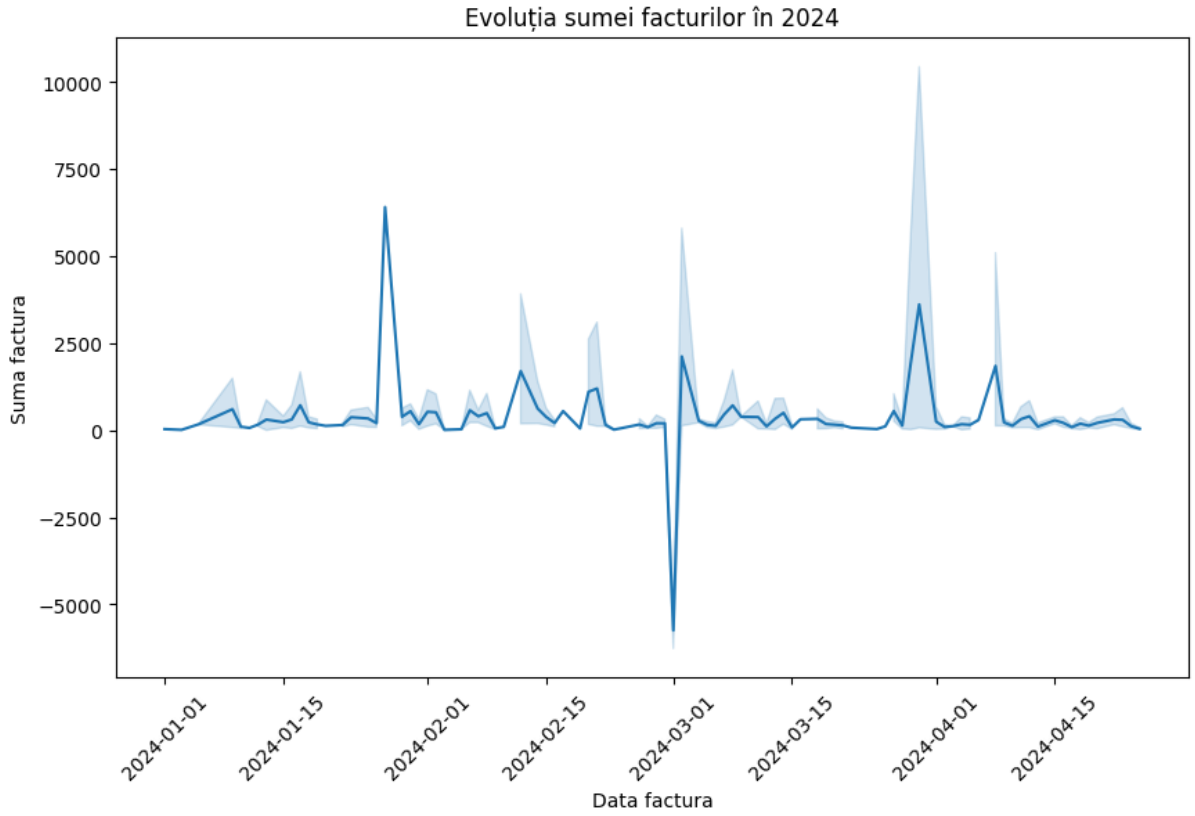

Evoluția sumei facturilor în 2019
Suma factura
0
500
1000
1500
2000
2500
3000
2019-01
2019-03
2019-05
2019-07
2019-09
2019-11
2020-01
Data factura


Evoluția sumei facturilor în 2020
Suma factura
0
2000
4000
6000
8000
2020-01
2020-03
2020-05
2020-07
2020-09
2020-11
2021-01
Data factura

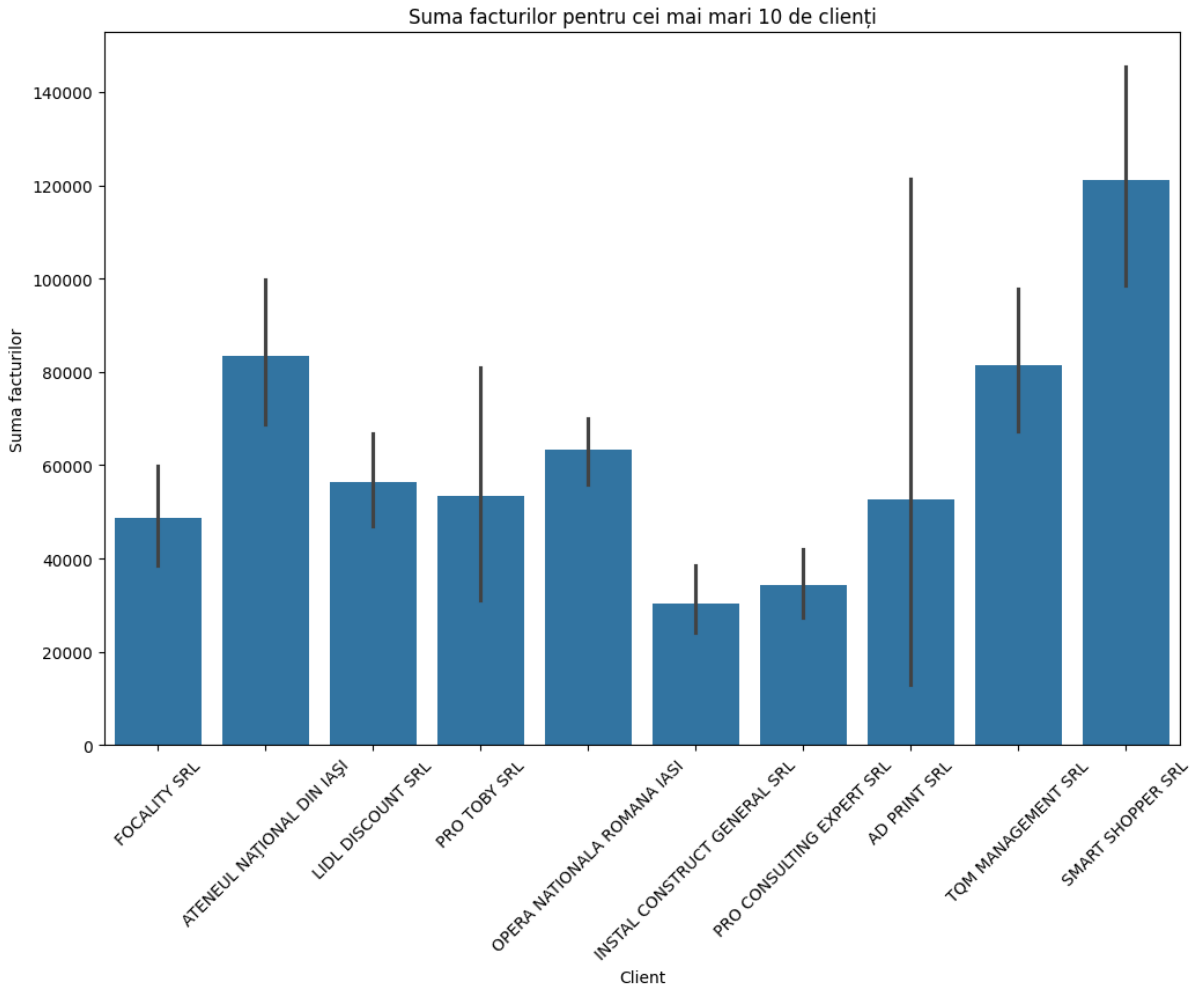

Evoluția sumei facturilor în 2021
Suma factura
0
5000
10000
15000
20000
2021-01
2021-03
2021-05
2021-07
2021-09
2021-11
2022-01
Data factura


Evoluția sumei facturilor în 2022
Suma factura
10000
8000
6000
4000
2000
0
2022-01
2022-03
2022-05
2022-07
2022-09
2022-11
2023-01
Data factura


Evoluția sumei facturilor în 2023
Suma factura
8000
6000
4000
2000
0
2023-01
2023-03
2023-05
2023-07
2023-09
2023-11
2024-01
Data factura


Evoluția sumei facturilor în 2024
Suma factura
10000
7500
5000
2500
0
−2500
−5000
2024-01-01
2024-01-15
2024-02-01
2024-02-15
2024-03-01
2024-03-15
2024-04-01
2024-04-15
Data factura

Suma facturilor pentru top 10 clienti.



Partea de Machine Learning – Antrenare Model:

```
# Exemplu de pregătire a datelor pentru regresie liniară
X = df[['Baza impozitare', 'Valoare TVA']]
y = df['Suma factura']
X_train, X_test, y_train, y_test = train_test_split(X, y,
test_size=0.2, random_state=42)

model = LinearRegression()
model.fit(X_train, y_train)
y_pred = model.predict(X_test)

# Evaluate model
mse = mean_squared_error(y_test, y_pred)
r2 = r2_score(y_test, y_pred)
print("Mean Squared Error:", mse)
print("R^2 Score:", r2)
```

Rezultatul obținut:

```
Mean Squared Error: 3.652068269258797e-05
R^2 Score: 0.9999999999353562
```

Interpretare:

1. **Mean Squared Error (MSE)**: Acesta este o măsură a diferenței medii pătratice între valorile prezise de model și valorile reale. Cu cât MSE este mai mic, cu atât modelul prezice mai bine datele. În cazul meu, valoarea MSE este foarte mică (3.652068269258797e-05), ceea ce indică faptul că modelul prezice foarte bine valorile.

2. **R-squared Score (R^2)**: Acesta este o măsură a cât de bine variabilitatea în datele de intrare este explicată de model. R-squared variază între 0 și 1, unde 1 înseamnă că modelul explică perfect variația datelor și 0 înseamnă că modelul nu explică deloc variația datelor. În cazul meu, valoarea R^2 este aproape de 1 (0.9999999999353562), ceea ce indică faptul că modelul explică foarte bine variația datelor.

Predicția:

```
# Folosesim modelul pentru a face predicții pentru cererea viitoare
cerere_viitoare_pred = model.predict(df[['Baza impozitare', 'Valoare
TVA']])

# Afișează predicțiile pentru cererea viitoare
print(cerere_viitoare_pred)
```

Rezultatul obținut:

```
# Selecția primelor 50 de valori prezise și cele reale
primele_50_valori = df[['Predictie', 'Suma factura']].head(50)

# Afișarea într-un tabel
print(primele_50_valori.to_string(index=False))
```

Predictie	Suma facturi reală
30.246328	30.24
407.514988	407.51
56.006236	56.00
270.445474	270.44
180.005795	180.00
51.106253	51.10
45.236274	45.23
1429.501357	1429.50
371.345116	371.34
94.006101	94.00
12.306391	12.30
195.305741	195.30
32.856318	32.85
8082.757724	8082.78
306.305347	306.30

57.006233	57.00
771.003696	771.00
62.006215	62.00
214.005675	214.00
245.205564	245.20
720.003878	720.00
55.806237	55.80
34.026314	34.02
598.254310	598.25
26.606340	26.60
58.686226	58.68
216.005667	216.00
225.005636	225.00
321.265294	321.26
4.806418	4.80
23.046353	23.04
15.126381	15.12
193.505748	193.50
151.155898	151.15
124.505992	124.50
5.766415	5.76
92.906105	92.90
491.004691	491.00
308.165340	308.16
74.906169	74.90
2542.947402	2542.95
311.345329	311.34
1200.522170	1200.52
537.004527	537.00
19.326366	19.32
147.005913	147.00
232.835608	232.83
157.485875	157.48
84.606135	84.60
797.993600	797.99

Există diferențe între valorile prezise și cele reale: