Setul de date

Am utilizat de pe kaggle urmatorul set de date:
https://www.kaggle.com/datasets/sevgisarac/temperature-change

Analiza exploratorie a datelor:

In acest set de date sunt 4 fisiere .csv.

In primul fisier sunt datele care acoperă perioada 1961–2019 și includ anomalii de temperatură lunare, sezoniere și anuale, adică schimbările de temperatură în raport cu o climatologie de bază, corespunzătoare perioadei 1951–1980. De asemenea, sunt disponibile statistici despre abaterea standard a schimbării temperaturii conform metodologiei de bază. Datele sunt bazate pe datele GISTEMP publice, distribuite de NASA-GISS (Institutul de Studii al Temperaturii la Suprafață Globală). Pentru a putea procesa si vizualiza datele in continuare am pastrat doar coloanele numerice: **Area Code, Months Code, Element Code, Year si Valoarea respectiva.**

Din celelalte fisiere, cel pentru 2022 si 2024 am extras aceleasi coloane. In 2024 Area Code era in format M49 si pentru a corespunde cu ce aveam in celelalte doua fisiere, am inlocuit formatul M$8 cu formatul FAO corespunzator folosind fisierul din 2020.

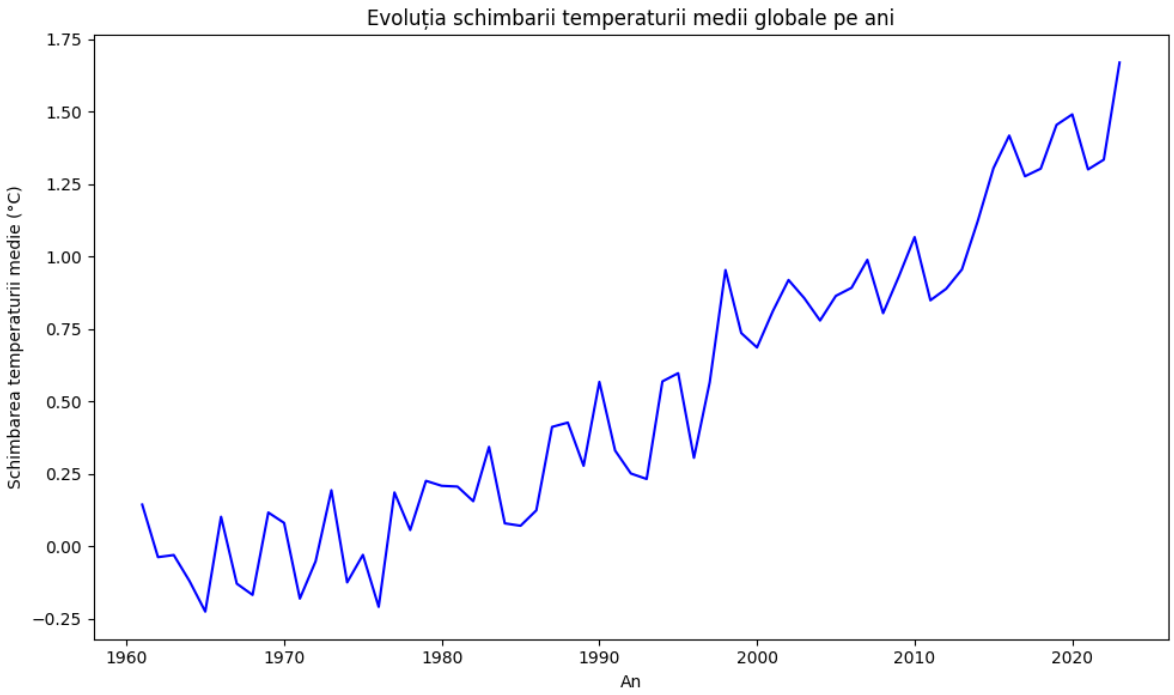
Rezultatul functiei describe dupa ce am elimiant coloanele de tip text si am combinat cele 3 csv-uri:

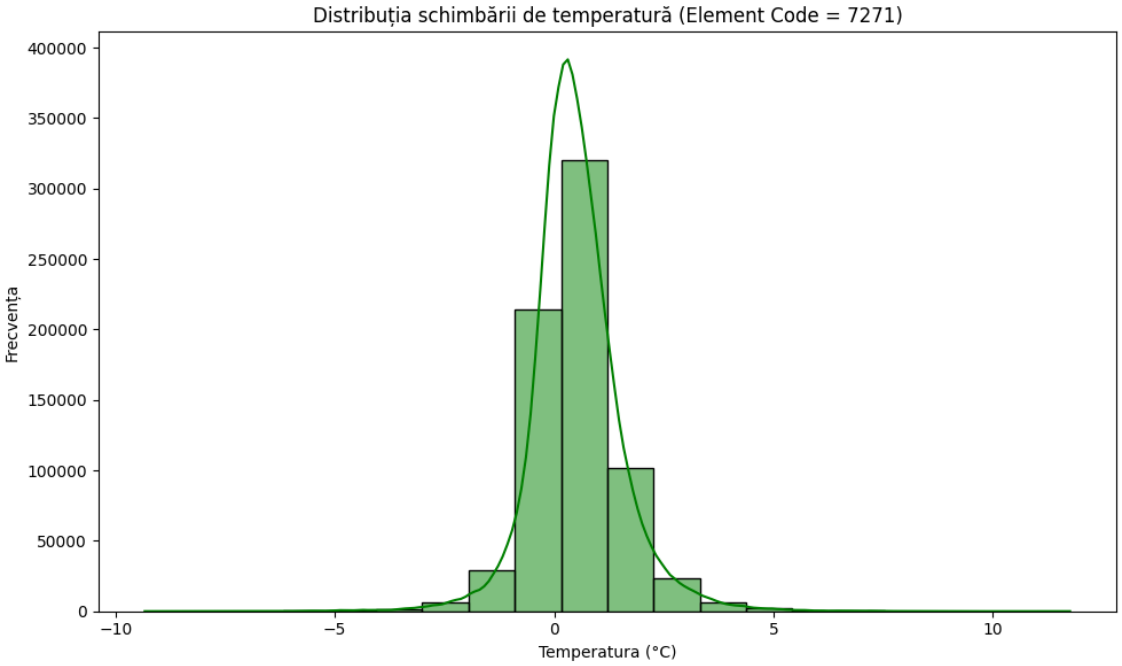
	Area Code	Months Code	Element Code	Year	Value
count	1.040451e+06	1.040451e+06	1.040451e+06	1.040451e+06	944175.000000
mean	5.088993e+02	7.009882e+03	6.944384e+03	1.990944e+03	0.544909
std	1.363050e+03	6.037945e+00	5.319543e+02	1.740521e+01	0.925293
min	1.000000e+00	7.001000e+03	6.078000e+03	1.961000e+03	-9.334000
25%	7.000000e+01	7.005000e+03	6.078000e+03	1.976000e+03	0.113000
50%	1.410000e+02	7.009000e+03	7.271000e+03	1.991000e+03	0.465000
75%	2.110000e+02	7.016000e+03	7.271000e+03	2.006000e+03	0.956000
max	5.873000e+03	7.020000e+03	7.271000e+03	2.023000e+03	11.759000

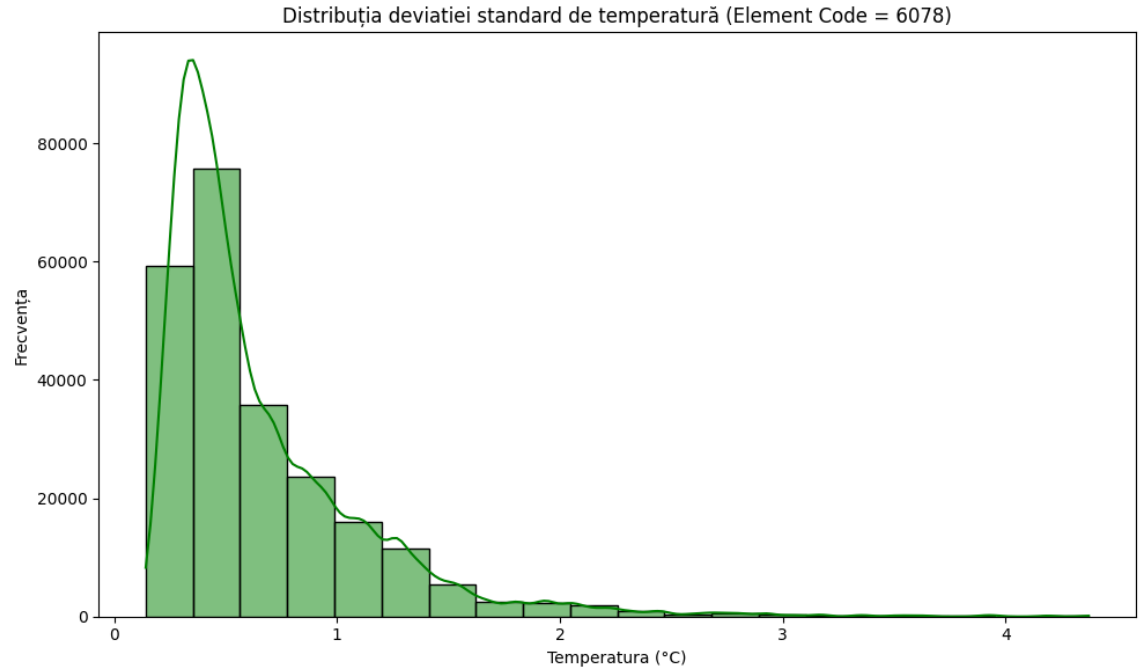
Valorile de tip null care ulterior au fost eliminate:

```
Area Code           0
Months Code         0
Element Code        0
Year                0
Value           96276
dtype: int64
Procentul de valori NaN în 'Value': 9.25%
```

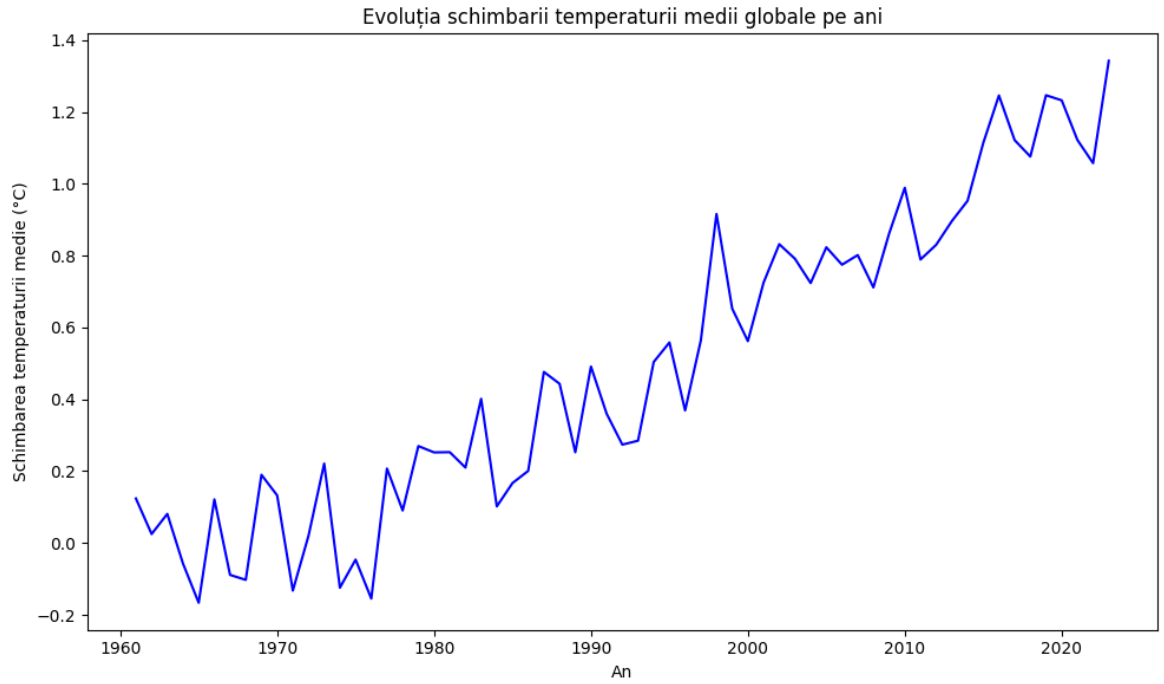
Grafice relevante:

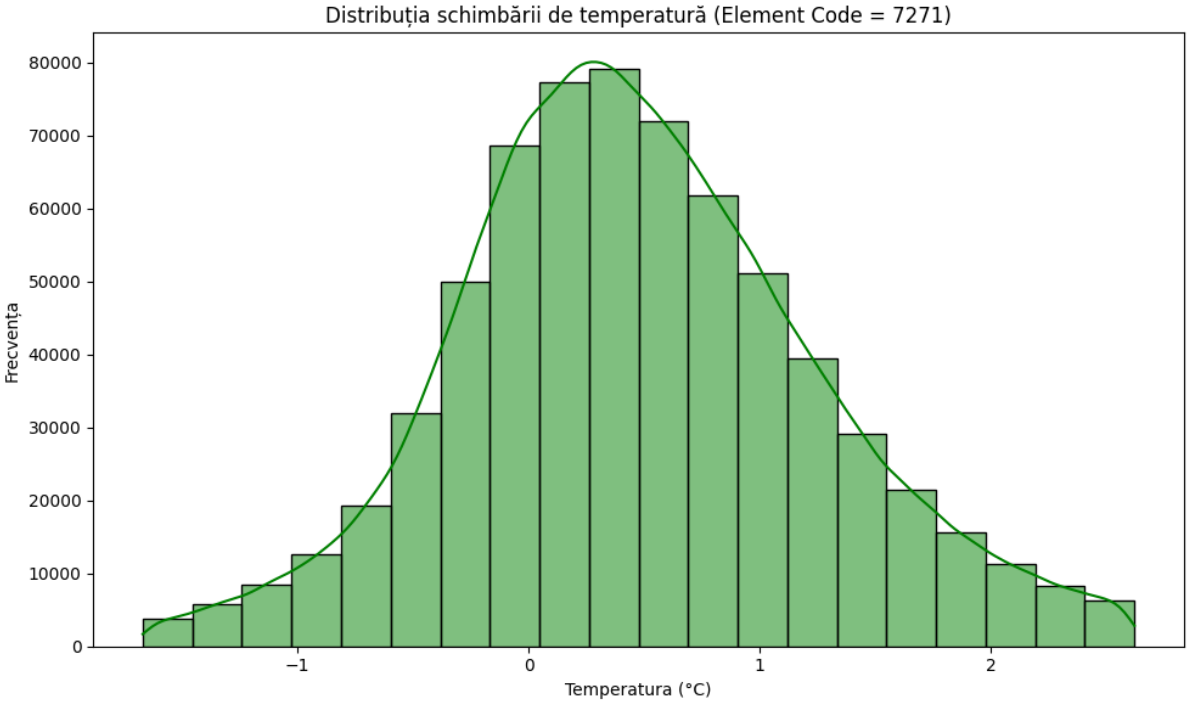


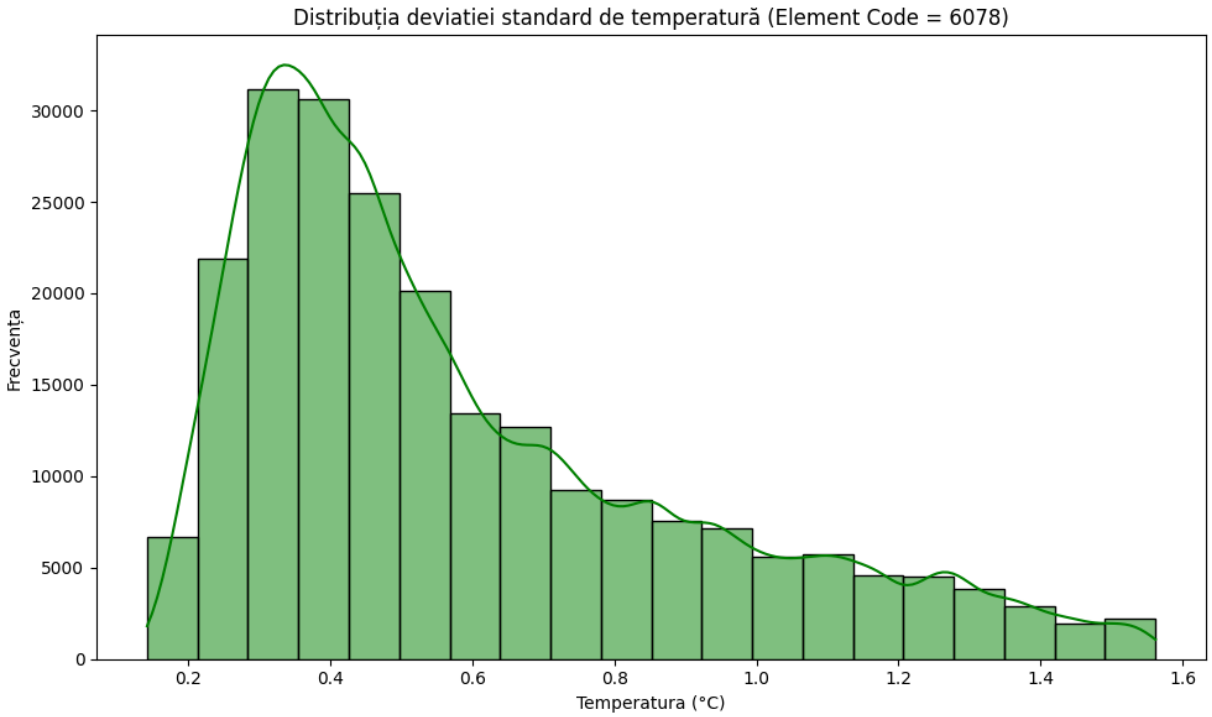




Aceste grafice sunt facute inainte de eliminarea outliers.


Evoluția schimbarii temperaturii medii globale pe ani
Schimbarea temperaturii medie (°C)
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
1960
1970
1980
1990
2000
2010
2020
An


Distribuția schimbării de temperatură (Element Code = 7271)
Frecvența
80000
70000
60000
50000
40000
30000
20000
10000
0
−1
0
1
2
Temperatura (°C)



Graficele dupa eliminarea outliers

Preprocesarea setului de date:

Setul de date conține două tipuri de informații: schimbarea de temperatură și deviația standard. Din acest motiv, am preprocesat fiecare dintre ele separat, pentru a asigura o gestionare corectă a fiecărui tip de dată.

Standardizare: Am aplicat **RobustScaler** pentru standardizarea datelor, deoarece această metodă este mai puțin sensibilă la valori extreme (outliers) și este utilă atunci când datele sunt distribuite neuniform sau conțin astfel de valori. Astfel, valorile au fost scalate pentru a avea o medie de 0 și o abatere standard de 1.

Codificare pentru caracteristici categorice: Am aplicat label encoding pentru a transforma valorile categorice din coloanele Months și Area Code în valori numerice. Aceasta a fost o alegere potrivită pentru aceste variabile, deoarece acestea sunt variabile de tipul „categoric ordinal” (există o ordine implicită în luna anului și în zonele geografice, dar nu este utilă o relație matematică între valorile lor).