ODREĐIVANJE I FILTRACIJA POŽELJNIH PODATAKA

Temperatura:

Da bismo odredili poželjne podatke (podatke koji su nam korisni za kasnije analiziranje) moramo izdvojiti i maknuti one koji iz nekog razloga previše odstupaju od ostalih. Da bismo to postigli, koristit ćemo 3-sigma pravilo. 3-sigma pravilo (pravilo 68-95-99.7) odnosi se na normalnu raspodjelu i kaže da se oko 68% podataka nalazi unutar 1 standardne devijacije, 95% unutar 2 standardne devijacije, i 99.7% unutar 3 standardne devijacije od srednje vrijednosti.

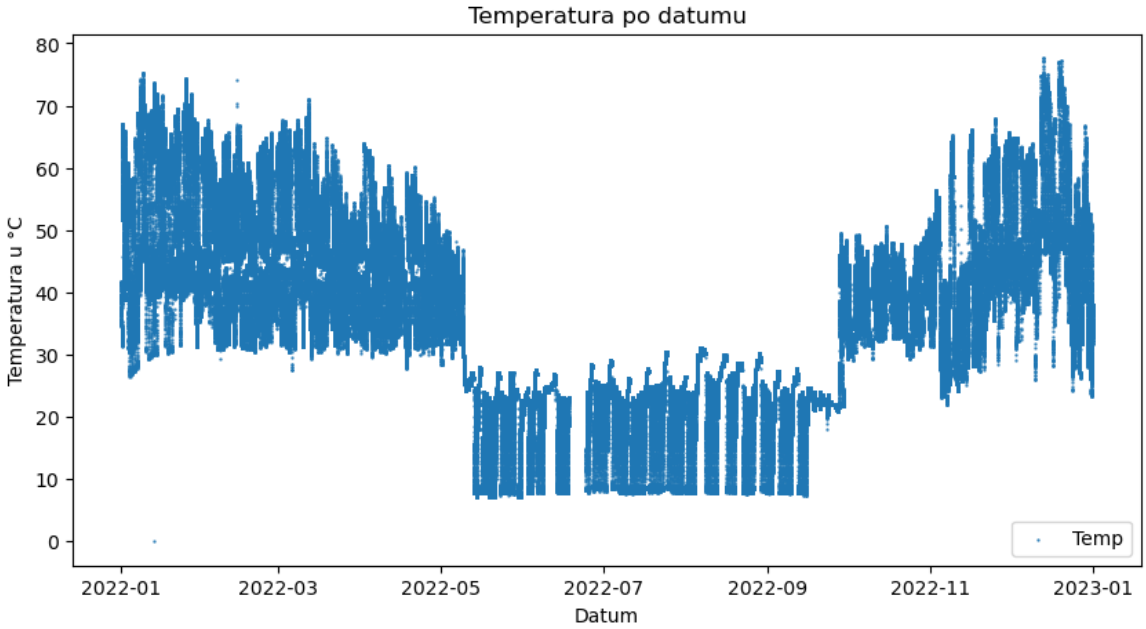
Kako bismo kvalitetno odredili grupiranje podataka, moramo odvojiti periode grijanja odnosno hlađenja (održavanja temperature). Da uzmemo prosjek cijele godine dobili bismo nelogične podatke jer on sadrži periode i grijanja i hlađenja. Iz prikazanih podataka vidimo da se negdje početkom svibnja događa promjena grijanje-hlađenje te krajem rujna hlađenje-grijanje pa ćemo u tim vremenskim periodima provesti detaljnije istraživanje i analizu podataka.

Za primjer ćemo uzeti 2022. godinu

In [10]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
kal22 = pd.read_csv('/Users/mateotoic/Desktop/ProjektR/Godina/2022/
```

In [12]:
```
kalcopy = kal22.copy()
kalcopy['timestamp'] = pd.to_datetime(kalcopy['timestamp'])

kal_sort = kalcopy.sort_values(by='timestamp')
```

In [14]:
```
plt.figure(figsize=(10,5))
plt.scatter(kal_sort.timestamp, kal_sort.supply_temperature, label=
plt.title("Temperatura po datumu")
plt.xlabel("Datum")
plt.ylabel("Temperatura u °C")
plt.legend()
plt.show()
```



Zaključujemo da je temperatura oko 30 stupnjeva Celzijevih granična (iznad nje se nalazi većina podataka kad grijemo prostorije, a ispod kad hladimo/održavamo temperaturu). Graf prosječne temperature nam samo potvrđuje hipotezu. Da bismo odredili točan datum promjene, promatramo u kojem se danu događa prijelaz kada većina temperatura više nije iznad 30 stupnjeva nego ispod.

In [16]:
```
kal_sort['interval_d'] = kal_sort['timestamp'].dt.to_period('D')

grupe_intervala = kal_sort.groupby('interval_d')

rezultat = []

for interval, group in grupe_intervala:
    ukupno = len(group)
    ispod_30 = len(group[group['supply_temperature'] < 30])
    iznad_30 = ukupno - ispod_30

    # Odredimo većinu
    vecina = "Ispod 30°C" if ispod_30 > iznad_30 else "Iznad 30°C"

    # Dodamo rezultate u listu
    rezultat.append({
        'interval': interval,
        'ukupno_podataka': ukupno,
        'ispod_30': ispod_30,
        'iznad_30': iznad_30,
        'vecina': vecina
    })

df_intervals = pd.DataFrame(rezultat)

konacni = df_intervals.query('interval > "2022-05-01" and interval
konacni
```

Out[16]:

	interval	ukupno_podataka	ispod_30	iznad_30	vecina
121	2022-05-02	1440	0	1440	Iznad 30°C
122	2022-05-03	1440	0	1440	Iznad 30°C
123	2022-05-04	1440	0	1440	Iznad 30°C
124	2022-05-05	1440	13	1427	Iznad 30°C
125	2022-05-06	1440	0	1440	Iznad 30°C
126	2022-05-07	1440	0	1440	Iznad 30°C
127	2022-05-08	1440	0	1440	Iznad 30°C
128	2022-05-09	1440	587	853	Iznad 30°C
129	2022-05-10	1440	1440	0	Ispod 30°C
130	2022-05-11	1440	1440	0	Ispod 30°C
131	2022-05-12	1440	1440	0	Ispod 30°C
132	2022-05-13	1440	1440	0	Ispod 30°C
133	2022-05-14	1440	1440	0	Ispod 30°C

Možemo zaključiti da je 9. svibnja dan kad se prestaje grijati. Istu stvar učinimo i obrnuto - tražimo datum kad se kreće grijati:

In [18]:

```
kal_sort['interval_d'] = kal_sort['timestamp'].dt.to_period('D')

grupe_intervala = kal_sort.groupby('interval_d')

rezultat = []

for interval, group in grupe_intervala:
    ukupno = len(group)
    ispod_30 = len(group[group['supply_temperature'] < 30])
    iznad_30 = ukupno - ispod_30

    # Odredimo većinu
    vecina = "Ispod 30°C" if ispod_30 > iznad_30 else "Iznad 30°C"

    # Dodamo rezultate u listu
    rezultat.append({
        'interval': interval,
        'ukupno_podataka': ukupno,
        'ispod_30': ispod_30,
        'iznad_30': iznad_30,
        'vecina': vecina
    })

df_intervals = pd.DataFrame(rezultat)

konacni = df_intervals.query('interval > "2022-09-15" and interval
konacni
```

Out[18]:

	interval	ukupno_podataka	ispod_30	iznad_30	vecina
253	2022-09-16	1440	1440	0	Ispod 30°C
254	2022-09-17	1440	1440	0	Ispod 30°C
255	2022-09-18	1440	1440	0	Ispod 30°C
256	2022-09-19	1440	1440	0	Ispod 30°C
257	2022-09-20	1440	1440	0	Ispod 30°C
258	2022-09-21	1440	1440	0	Ispod 30°C
259	2022-09-22	1440	1440	0	Ispod 30°C
260	2022-09-23	1440	1440	0	Ispod 30°C
261	2022-09-24	1440	1440	0	Ispod 30°C
262	2022-09-25	1440	1440	0	Ispod 30°C
263	2022-09-26	1440	1440	0	Ispod 30°C
264	2022-09-27	1440	1010	430	Ispod 30°C
265	2022-09-28	1440	11	1429	Iznad 30°C
266	2022-09-29	1334	361	973	Iznad 30°C
267	2022-09-30	1440	0	1440	Iznad 30°C

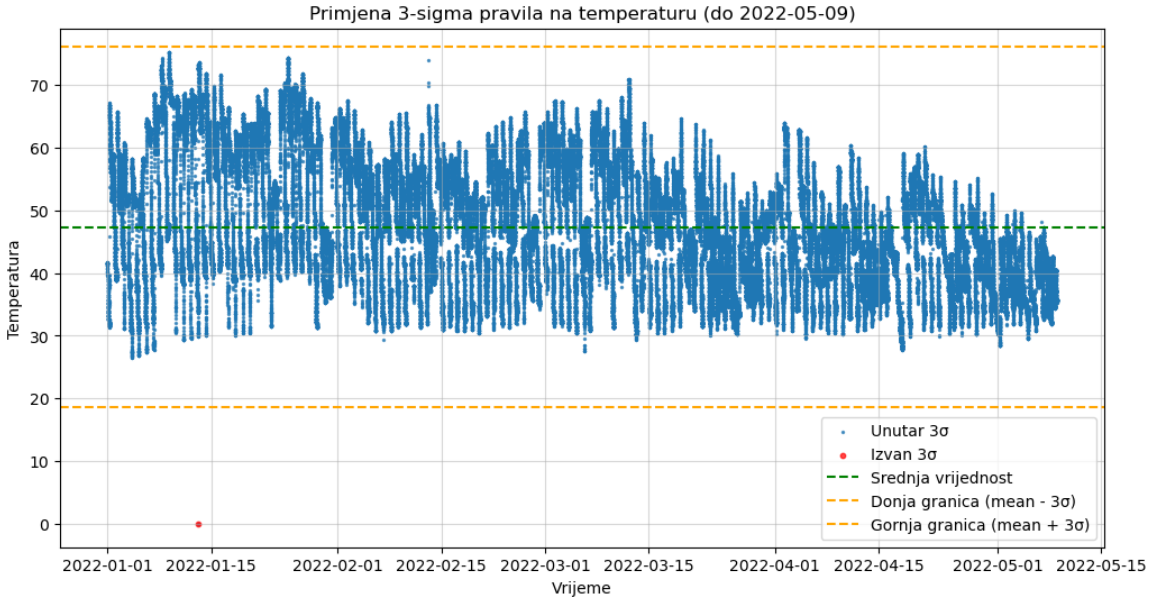
Vidimo da je traženi datum 28. rujna te sad kad imamo oba datuma granica perioda grijanja odnosno hlađenja možemo nastaviti s analizom.

Prvo iskoristimo 3-sigma pravilo nad prvim skupom podataka (1.1.2022. - 9.5.2022.)

In [20]:
```
filtrirani = kal_sort[kal_sort['timestamp'] <= "2022-05-09"]


mean_temp = filtrirani['supply_temperature'].mean()
std_temp = filtrirani['supply_temperature'].std()

donja_granica = mean_temp - 3 * std_temp
gornja_granica = mean_temp + 3 * std_temp

prihvatljivi = filtrirani[(filtrirani['supply_temperature'] >= donj
                                    (filtrirani['supply_temperature']

odbaceni = filtrirani[(filtrirani['supply_temperature'] < donja_gra
                                    (filtrirani['supply_temperature']

print(f"Srednja vrijednost: {mean_temp}")
print(f"Standardna devijacija: {std_temp}")
print(f"Interval (mean ± 3σ): [{donja_granica}, {gornja_granica}]")
print(f"Broj podataka unutar intervala: {len(filtrirani)}")
```

```
print(f"Broj outliera: {len(odbaceni)}")

plt.figure(figsize=(12,6))
plt.scatter(filtrirani['timestamp'], filtrirani['supply_temperature
plt.scatter(odbaceni['timestamp'], odbaceni['supply_temperature'],

plt.axhline(mean_temp, color='green', linestyle='--', label="Srednj
plt.axhline(donja_granica, color='orange', linestyle='--', label="D
plt.axhline(gornja_granica, color='orange', linestyle='--', label="
plt.title("Primjena 3-sigma pravila na temperaturu (do 2022-05-09)"
plt.xlabel("Vrijeme")
plt.ylabel("Temperatura")
plt.legend()
plt.grid(alpha=0.5)

plt.show()
```

```
Srednja vrijednost: 47.34353822078289
Standardna devijacija: 9.58448611223744
Interval (mean ± 3σ): [18.59007988407057, 76.0969965574952]
Broj podataka unutar intervala: 184036
Broj outliera: 1
```

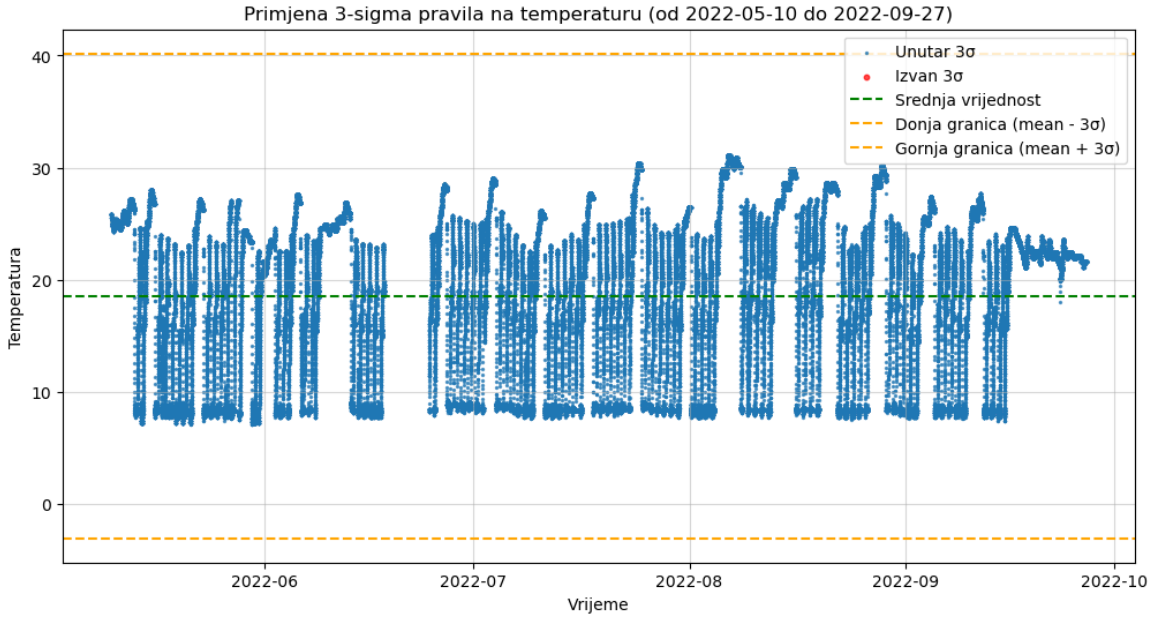


Zatim isto napravimo nad idućim skupovima:

In [23]:
```
filtrirani = kal_sort[(kal_sort['timestamp'] > "2022-05-10") & (kal

mean_temp = filtrirani['supply_temperature'].mean()
std_temp = filtrirani['supply_temperature'].std()

donja_granica = mean_temp - 3 * std_temp
gornja_granica = mean_temp + 3 * std_temp

prihvatljivi = filtrirani[(filtrirani['supply_temperature'] >= donj
                          (filtrirani['supply_temperature']

odbaceni = filtrirani[(filtrirani['supply_temperature'] < donja_gra
                          (filtrirani['supply_temperature']
```

```
print(f"Srednja vrijednost: {mean_temp}")
print(f"Standardna devijacija: {std_temp}")
print(f"Interval (mean ± 3σ): [{donja_granica}, {gornja_granica}]")
print(f"Broj podataka unutar intervala: {len(filtrirani)}")
print(f"Broj outliera: {len(odbaceni)}")

plt.figure(figsize=(12,6))
plt.scatter(filtrirani['timestamp'], filtrirani['supply_temperature
plt.scatter(odbaceni['timestamp'], odbaceni['supply_temperature'],

plt.axhline(mean_temp, color='green', linestyle='--', label="Srednj
plt.axhline(donja_granica, color='orange', linestyle='--', label="D
plt.axhline(gornja_granica, color='orange', linestyle='--', label="
plt.title("Primjena 3-sigma pravila na temperaturu (od 2022-05-10 d
plt.xlabel("Vrijeme")
plt.ylabel("Temperatura")
plt.legend()
plt.grid(alpha=0.5)

plt.show()
```

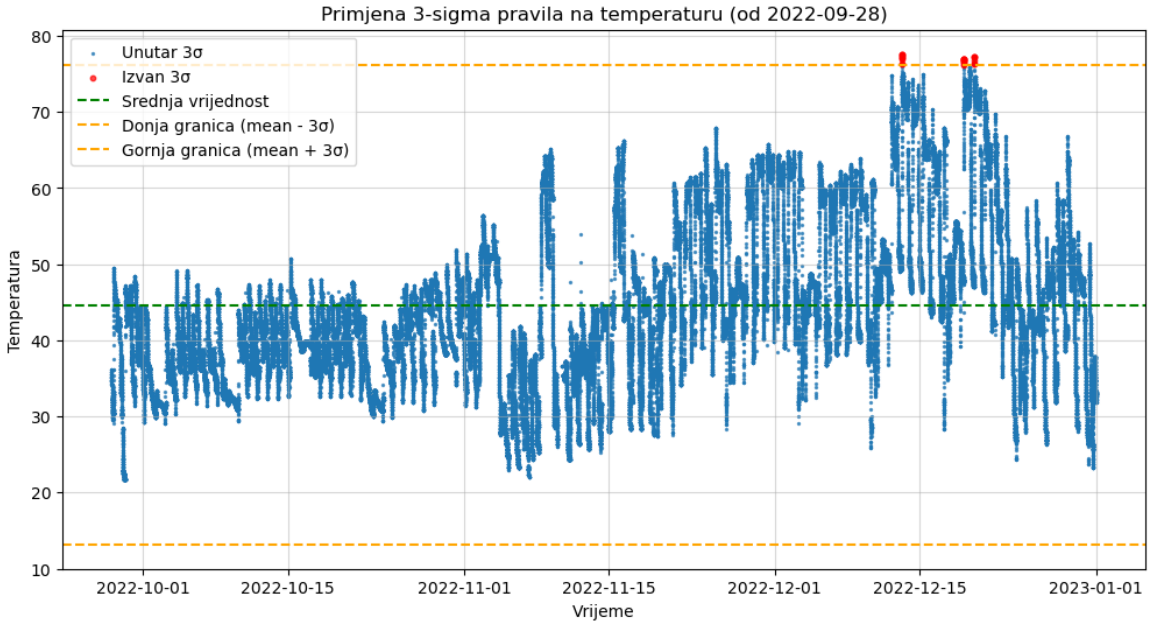
```
Srednja vrijednost: 18.567663706298074
Standardna devijacija: 7.2046870470000774
Interval (mean ± 3σ): [-3.0463974347021576, 40.181724847298305]
Broj podataka unutar intervala: 191868
Broj outliera: 0
```



In [26]:
```
filtrirani = kal_sort[kal_sort['timestamp'] >= "2022-09-28"]

mean_temp = filtrirani['supply_temperature'].mean()
std_temp = filtrirani['supply_temperature'].std()

donja_granica = mean_temp - 3 * std_temp
gornja_granica = mean_temp + 3 * std_temp

prihvatljivi = filtrirani[(filtrirani['supply_temperature'] >= donj
                          (filtrirani['supply_temperature']

odbaceni = filtrirani[(filtrirani['supply_temperature'] < donja_gra
```

```
                                        (filtrirani['supply_temperature']

print(f"Srednja vrijednost: {mean_temp}")
print(f"Standardna devijacija: {std_temp}")
print(f"Interval (mean ± 3σ): [{donja_granica}, {gornja_granica}]")
print(f"Broj podataka unutar intervala: {len(filtrirani)}")
print(f"Broj outliera: {len(odbaceni)}")

plt.figure(figsize=(12,6))
plt.scatter(filtrirani['timestamp'], filtrirani['supply_temperature
plt.scatter(odbaceni['timestamp'], odbaceni['supply_temperature'],

plt.axhline(mean_temp, color='green', linestyle='--', label="Srednj
plt.axhline(donja_granica, color='orange', linestyle='--', label="D
plt.axhline(gornja_granica, color='orange', linestyle='--', label="
plt.title("Primjena 3-sigma pravila na temperaturu (od 2022-09-28)"
plt.xlabel("Vrijeme")
plt.ylabel("Temperatura")
plt.legend()
plt.grid(alpha=0.5)

plt.show()
```

```
Srednja vrijednost: 44.66305596622461
Standardna devijacija: 10.488606695377927
Interval (mean ± 3σ): [13.19723588009083, 76.1288760523584]
Broj podataka unutar intervala: 135957
Broj outliera: 25
```

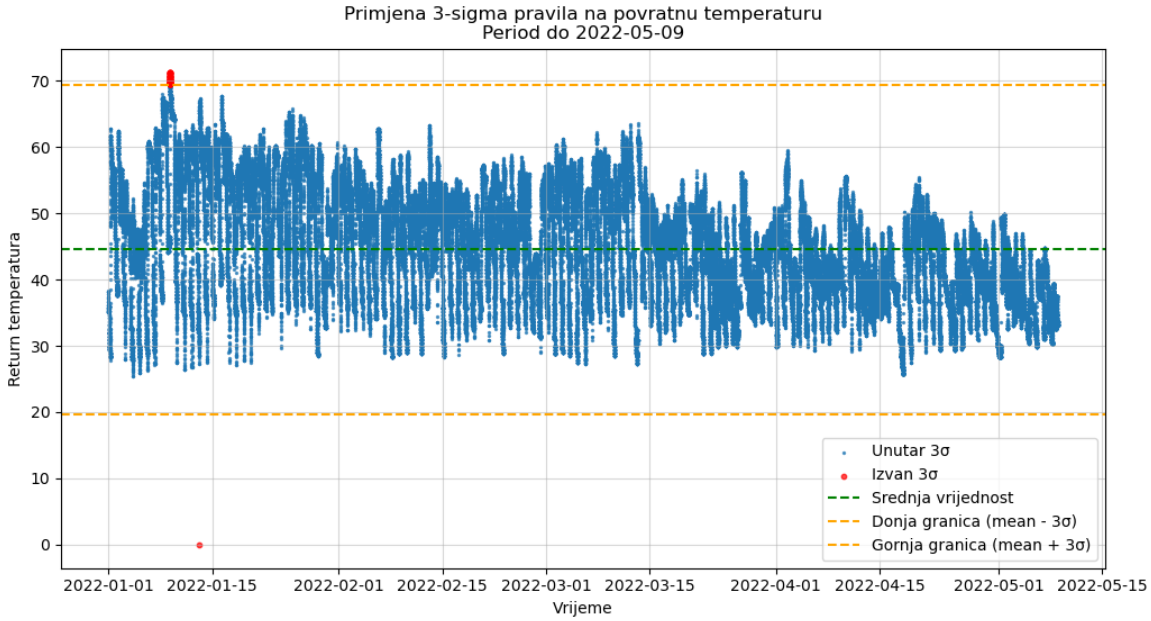


Isto napravimo i za 'return_temperature' te zatim i za 'mass_volume_flow'.

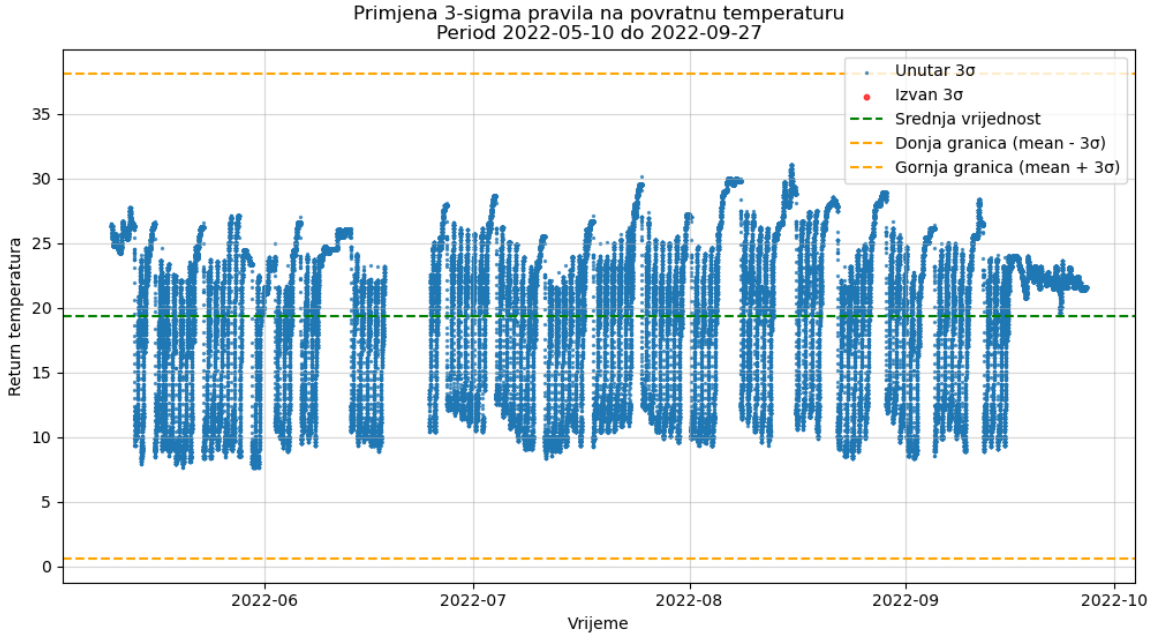
In [55]:
```
intervali = [
    ("Period do 2022-05-09", kal_sort[kal_sort['timestamp'] <= "202
    ("Period 2022-05-10 do 2022-09-27",
     kal_sort[(kal_sort['timestamp'] > "2022-05-10") & (kal_sort['t
    ("Period od 2022-09-28", kal_sort[kal_sort['timestamp'] >= "202
]
```

```
for naziv_intervala, filtrirani in intervali:

    mean_temp = filtrirani['return_temperature'].mean()
    std_temp = filtrirani['return_temperature'].std()

    donja_granica = mean_temp - 3 * std_temp
    gornja_granica = mean_temp + 3 * std_temp

    prihvatljivi = filtrirani[
        (filtrirani['return_temperature'] >= donja_granica) &
        (filtrirani['return_temperature'] <= gornja_granica)
    ]
    odbaceni = filtrirani[
        (filtrirani['return_temperature'] < donja_granica) |
        (filtrirani['return_temperature'] > gornja_granica)
    ]



    print(f"Interval: {naziv_intervala}")
    print(f"Srednja vrijednost: {mean_temp}")
    print(f"Standardna devijacija: {std_temp}")
    print(f"Interval (mean ± 3σ): [{donja_granica}, {gornja_granica
    print(f"Broj podataka (filtrirani): {len(filtrirani)}")
    print(f"Broj outliera (izvan 3σ): {len(odbaceni)}\n")


    plt.figure(figsize=(12, 6))
    plt.scatter(prihvatljivi['timestamp'], prihvatljivi['return_tem
                label="Unutar 3σ", s=2, alpha=0.7)
    plt.scatter(odbaceni['timestamp'], odbaceni['return_temperature
                label="Izvan 3σ", s=10, color='red', alpha=0.7)

    plt.axhline(mean_temp, color='green', linestyle='--', label="Sr
    plt.axhline(donja_granica, color='orange', linestyle='--', labe
    plt.axhline(gornja_granica, color='orange', linestyle='--', lab

    plt.title(f"Primjena 3-sigma pravila na povratnu temperaturu\n{
    plt.xlabel("Vrijeme")
    plt.ylabel("Return temperatura")
    plt.legend()
    plt.grid(alpha=0.5)

    plt.show()
```

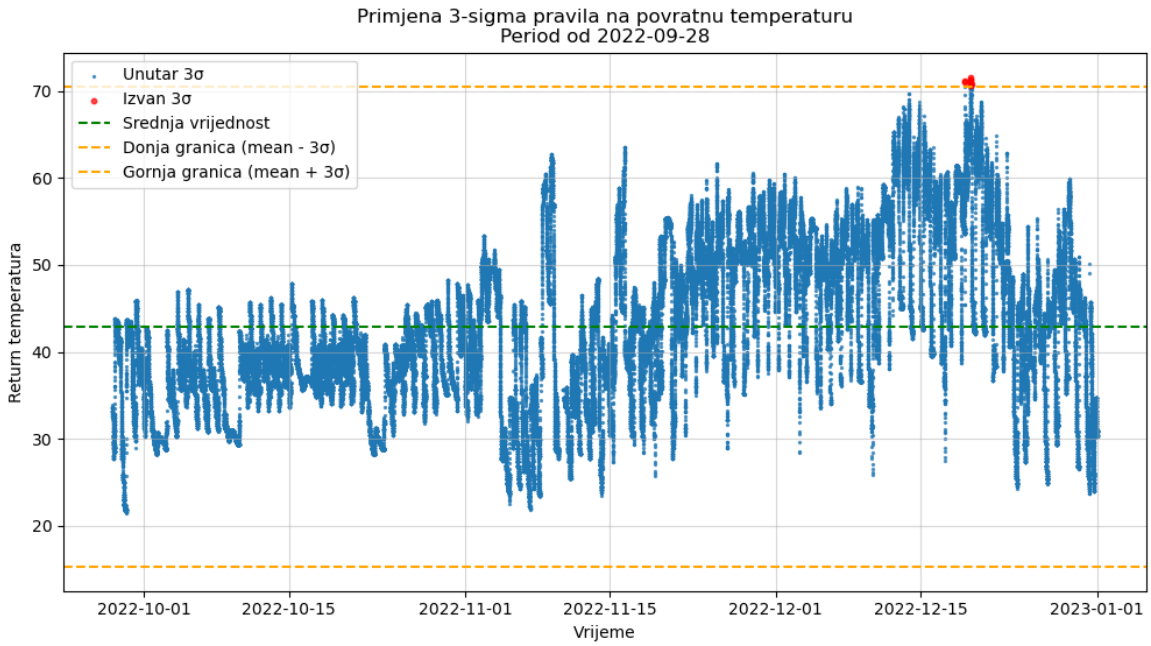
```
Interval: Period do 2022-05-09
Srednja vrijednost: 44.53499098002566
Standardna devijacija: 8.293342467646639
Interval (mean ± 3σ): [19.654963577085745, 69.41501838296557]
Broj podataka (filtrirani): 184036
Broj outliera (izvan 3σ): 45
```



Interval: Period 2022-05-10 do 2022-09-27
Srednja vrijednost: 19.374177038380548
Standardna devijacija: 6.244399934523619
Interval (mean ± 3σ): [0.6409772348096929, 38.10737684195141]
Broj podataka (filtrirani): 191868
Broj outliera (izvan 3σ): 0



Interval: Period od 2022-09-28
Srednja vrijednost: 42.934059298160456
Standardna devijacija: 9.19874022111907
Interval (mean ± 3σ): [15.337838634803244, 70.53027996151766]
Broj podataka (filtrirani): 135957
Broj outliera (izvan 3σ): 10

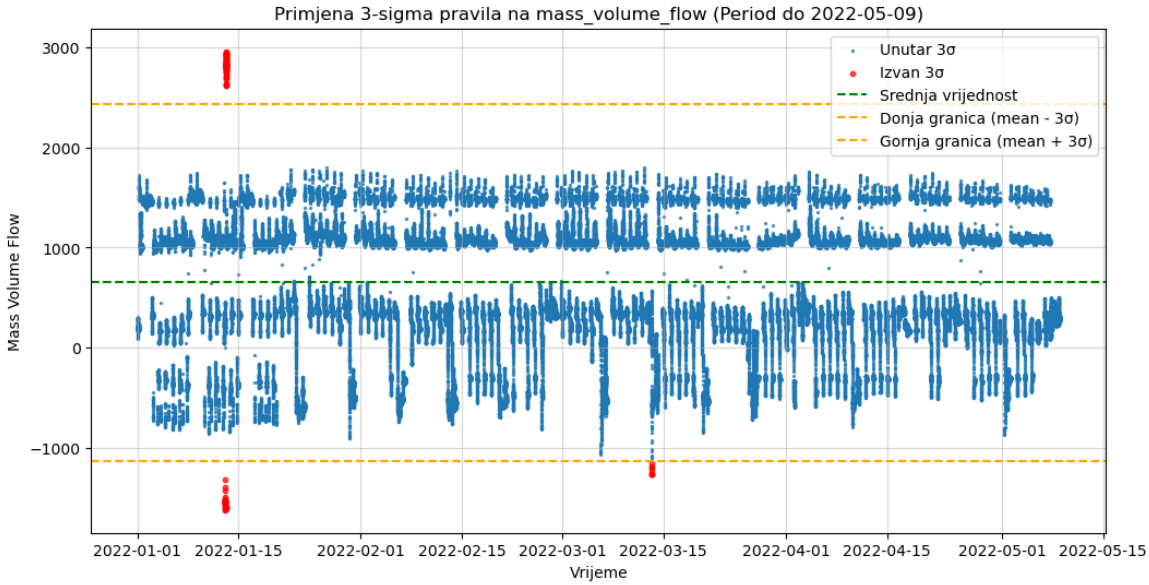


Maseni protok

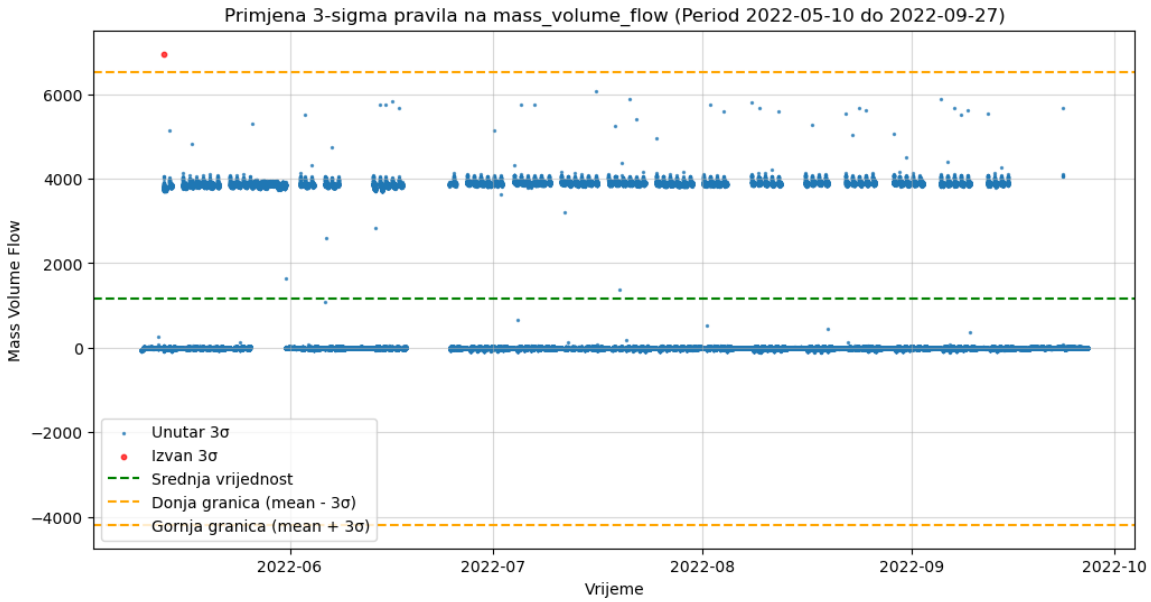
In [66]:
```
intervali = [
    ("Period do 2022-05-09", kal_sort[kal_sort['timestamp'] <= "202
    ("Period 2022-05-10 do 2022-09-27",
     kal_sort[(kal_sort['timestamp'] > "2022-05-10") & (kal_sort['t
    ("Period od 2022-09-28", kal_sort[kal_sort['timestamp'] >= "202
]

for naziv_intervala, filtrirani in intervali:

    mean_flow = filtrirani['mass_volume_flow'].mean()
    std_flow = filtrirani['mass_volume_flow'].std()


    donja_granica = mean_flow - 3 * std_flow
    gornja_granica = mean_flow + 3 * std_flow


    prihvatljivi = filtrirani[
        (filtrirani['mass_volume_flow'] >= donja_granica) &
        (filtrirani['mass_volume_flow'] <= gornja_granica)
    ]
    odbaceni = filtrirani[
        (filtrirani['mass_volume_flow'] < donja_granica) |
        (filtrirani['mass_volume_flow'] > gornja_granica)
    ]

    print(f"Interval: {naziv_intervala}")
    print(f"Srednja vrijednost protoka: {mean_flow}")
    print(f"Standardna devijacija: {std_flow}")
    print(f"Interval (mean ± 3σ): [{donja_granica}, {gornja_granica
    print(f"Broj podataka unutar intervala: {len(prihvatljivi)}")
    print(f"Broj outliera: {len(odbaceni)}\n")
```

```
    plt.figure(figsize=(12, 6))
    plt.scatter(prihvatljivi['timestamp'], prihvatljivi['mass_volum
                label="Unutar 3σ", s=2, alpha=0.7)
    plt.scatter(odbaceni['timestamp'], odbaceni['mass_volume_flow']
                label="Izvan 3σ", s=10, color='red', alpha=0.7)

    plt.axhline(mean_flow, color='green', linestyle='--', label="Sr
    plt.axhline(donja_granica, color='orange', linestyle='--', labe
    plt.axhline(gornja_granica, color='orange', linestyle='--', lab

    plt.title(f"Primjena 3-sigma pravila na mass_volume_flow ({nazi
    plt.xlabel("Vrijeme")
    plt.ylabel("Mass Volume Flow")
    plt.legend()
    plt.grid(alpha=0.5)

    plt.show()
```

```
Interval: Period do 2022-05-09
Srednja vrijednost protoka: 651.6653263491926
Standardna devijacija: 595.4037363988676
Interval (mean ± 3σ): [-1134.5458828474102, 2437.8765355457954]
Broj podataka unutar intervala: 183918
Broj outliera: 118
```




```
Interval: Period 2022-05-10 do 2022-09-27
Srednja vrijednost protoka: 1167.9711572539454
Standardna devijacija: 1784.2556482349305
Interval (mean ± 3σ): [-4184.795787450847, 6520.738101958737]
Broj podataka unutar intervala: 191867
Broj outliera: 1
```



Interval: Period od 2022-09-28
Srednja vrijednost protoka: 562.6427473392323
Standardna devijacija: 674.10469717882
Interval (mean ± 3σ): [-1459.6713441972274, 2584.956838875692]
Broj podataka unutar intervala: 135956
Broj outliera: 1