Beste Sanne,

Rosalinde vertelde dat je potentieel openstaat om ons iets te leren over het opschonen van data. Dat zou heel fijn zijn want tot nu toe hebben we dat een beetje op gevoel gedaan😊 We staan natuurlijk open voor alle soorten input/lessen/wijsheden, maar als opzet dachten we uit te leggen wat wij hebben gedaan zodat je ons zou kunnen vertellen hoe jij het anders/beter/sneller/slimmer zou doen!

**Wat algemene info over de data**

* Er zijn 46 kastjes die in principe elk kwartier een temperatuurmeting doen en die verzenden via het Lohra-netwerk
* In veel gevallen (~25%?) komt de meting niet door, het vermoeden is dat dit komt doordat de meting niet verzonden wordt, niet doordat het niet gedaan wordt. Dit verschilt erg per kastje.
* We hebben geen informatie over hoe het kastje hangt (oriëntatie, zon/schaduw, etc), wel over de locatie (gps-coördinaten).
* Ze zouden buiten moeten hangen, maar sommige metingen doen vermoeden dat er wel eens eentje binnen staat. Dit zien we aan constante temperaturen.
* We hebben het vermoeden dat sommige metingen heel hoog uitvallen doordat het kastje in de volle zon staat.

**Waarom opschonen**

* Sommige metingen zijn duidelijk veel te hoog en andere veel te laag
* Kastjes die binnen staan willen we niet meerekenen

**Wat maakt het lastig**

* Wanneer is een meting té hoog of té laag? Een spreiding in de temperatuur is natuurlijk te verwachten.

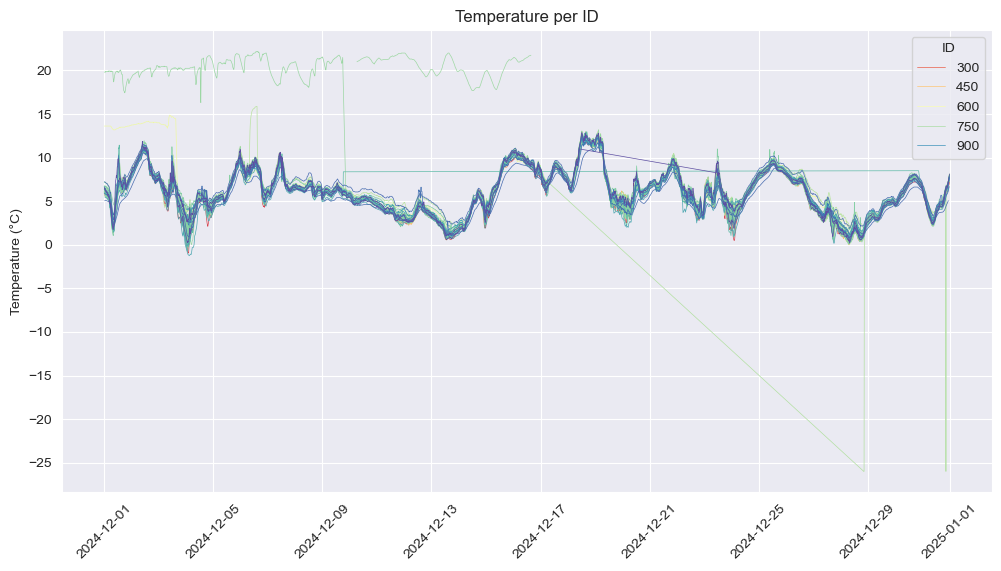
**Voorbeeld: data december**

Het opschonen hebben we gedaan door de data te plotten en de abnormaliteiten eruit te halen.

Vraag: is dit handig, of is een algemenere methode beter?

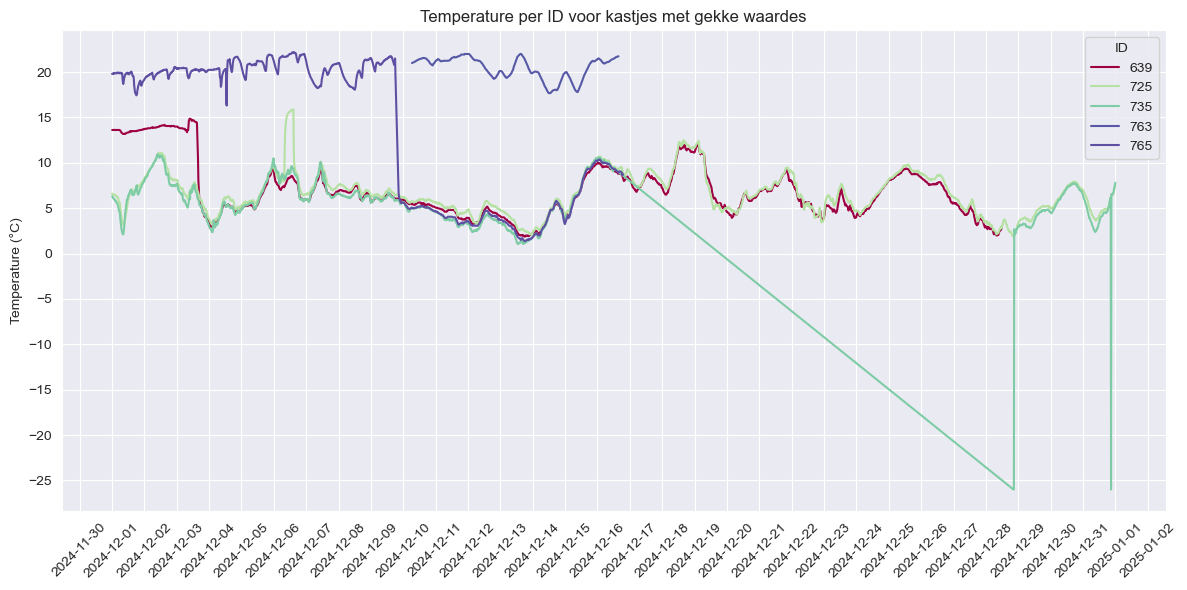
**Stappen**

1. Alle metingen per kastjes in een plot (afbeelding 1)



Afbeelding 1

1. Je ziet een aantal gekke metingen, die zijn apart geplot door alle kastjes eruit te halen met temperaturen boven de 14 en onder de -5 graden (afbeelding 2)

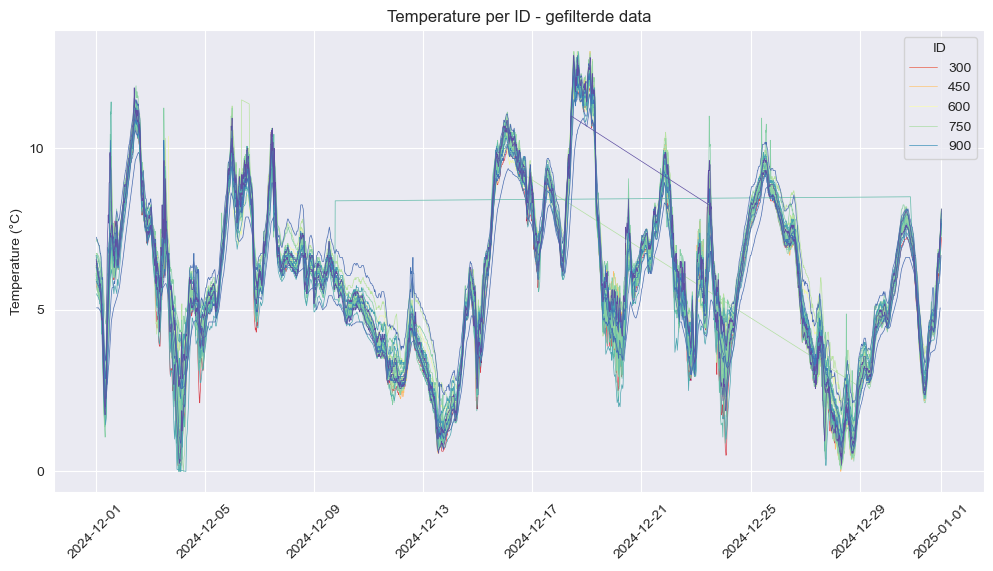


Afbeelding 2

1. Toen per kastje bepaald wat te doen:
   * + - 763 helemaal eruit
       - 765 t/m 9 december eruit
       - 639 t/m 3 december eruit
       - En verder alle losse metingen boven de 13 en onder de 0 graden eruit

Vraag: is dit een goede methode of is het beter om alle metingen van een bepaald kastje te verwijderen?

Dan krijg je een opgeschoonde plot (afbeelding 3), er zitten nog wat gekke metingen in, maar die heb ik laten zitten omdat ze (op het oog) hoogstens een paar graden buiten het gemiddelde vallen. Je zou hier nog kunnen verfijnen, maar dan heb je ook kans dat je echte metingen (bijvoorbeeld van kastjes in de zon) eruit haalt

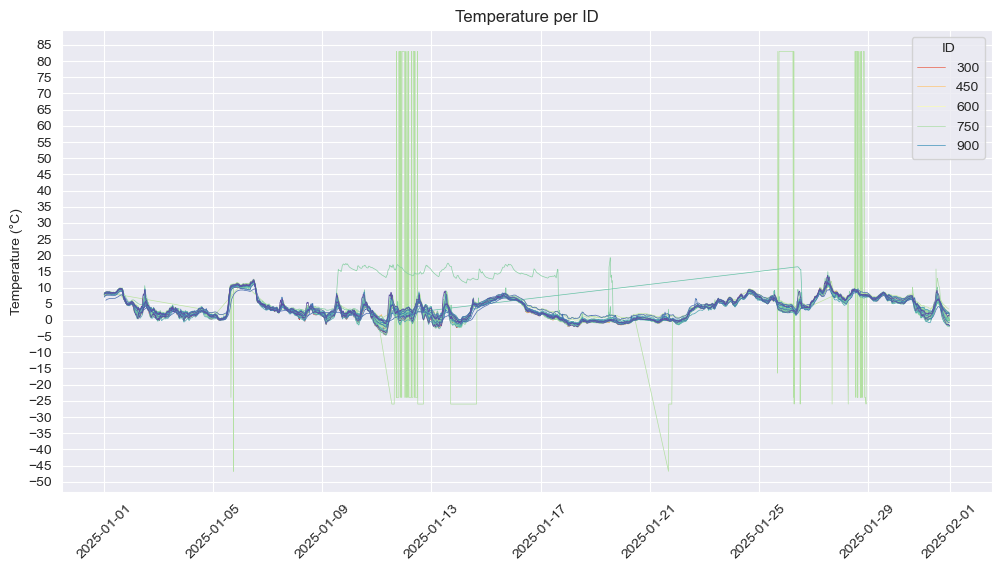


Afbeelding 3

**Voorbeeld: data januari**

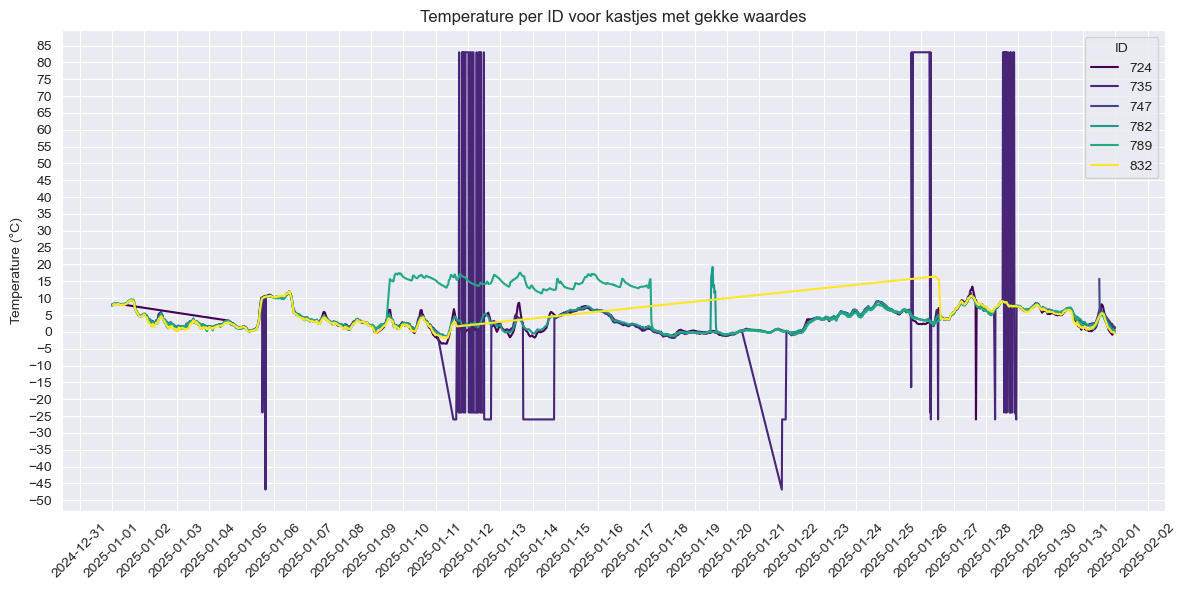
Zelfde methode als voor december

Ongefilterde data in afbeelding 4:



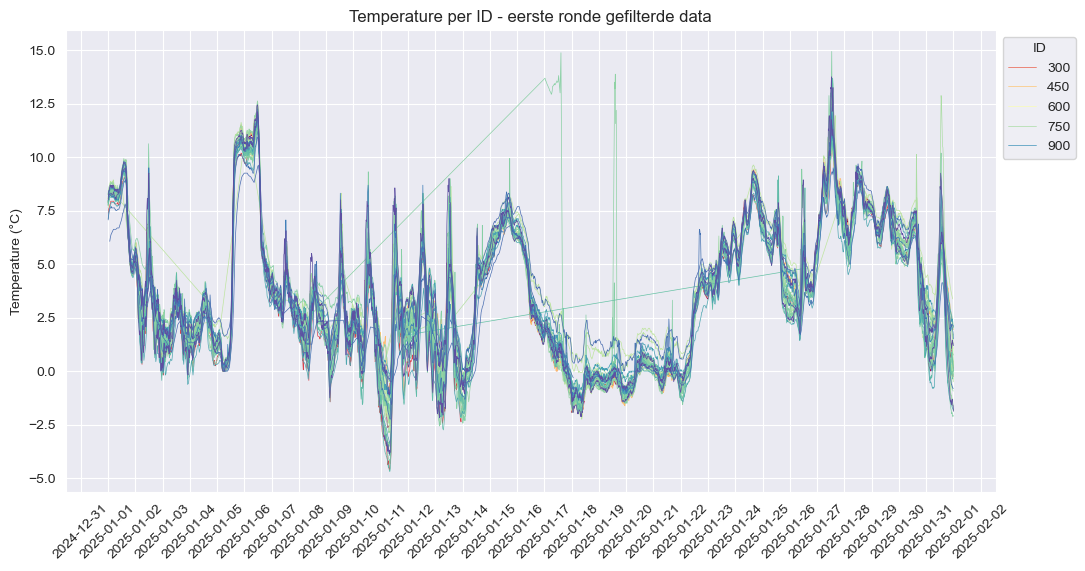
Afbeelding 4

Kastjes met temperaturen boven de 15 of onder de -10 geplot in afbeelding 5. Vervolgens alle individuele metingen boven de 15 en onder de -5 verwijderd. De rest lijkt namelijk vrij normaal.



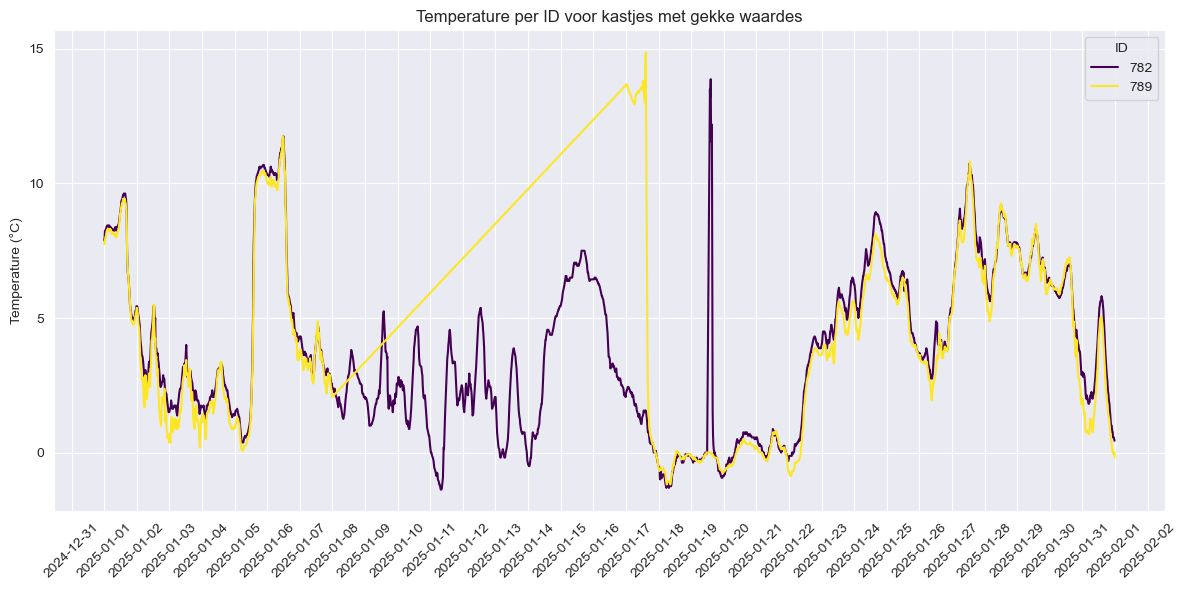
Afbeelding 5

Het resultaat is te zien in afbeelding 6. Er zit nog steeds een aantal uitschieters in.



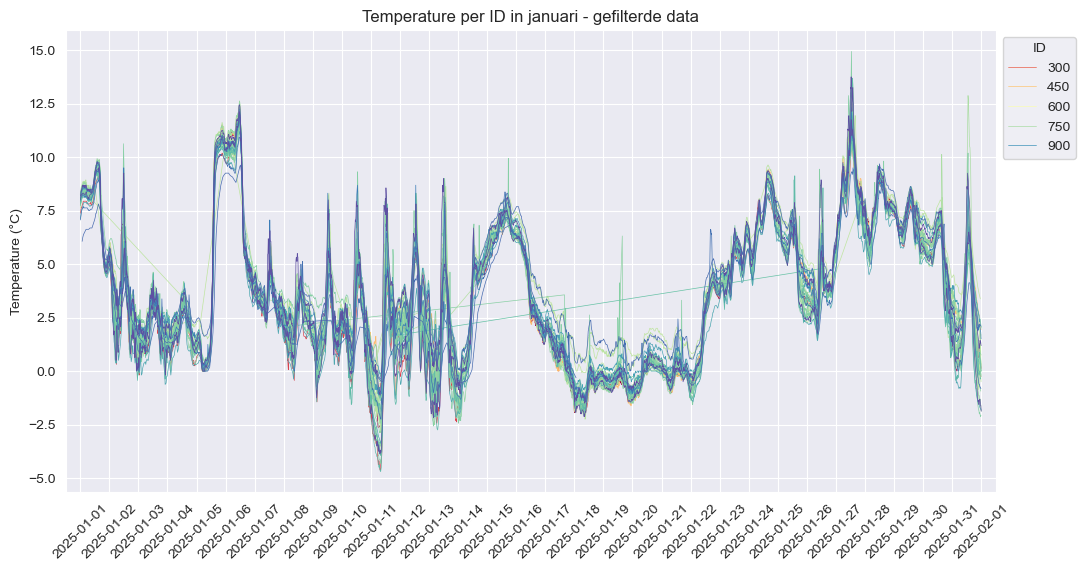
Afbeelding 6

De kastjes met de uitschieters geplot door alle kastjes te selecteren waar de temperatuur tussen 14 en 20 januari boven de 11 graden kwam. Zie afbeelding 7.



Afbeelding 7

Alle metingen tussen 14 en 20 januari boven de 11 graden verwijderd. Het eindresultaat is te zien in afbeelding 8. Wederom nog steeds niet perfect, op 18 januari zijn er bijvoorbeeld nog flinke uitschieters, maar de vraag is hoe ver we zouden moeten verfijnen? Is het bijvoorbeeld erg als we wat te veel data zouden verwijderen?



Afbeelding 8