# Final assignment programming 1

This project will answers two research questions mentioned below. For this project two datasets will be used from the National Institute for Health and Environment Netherlands.

The first dataset is the amount of corona RNA particles in sewage water on different locations in the provinces of the Netherlands.

The second dataset is the total amount of corona cases in each province of the Netherlands.

The sewage dataset has a timeline from 3/30/2020 to 1/27/2021. This timeline is shorter than that of the total amount dataset. Because of this the data will be filtered on this timeline to give a more representative result.

A third dataset is used for the mapping of the corona cases on the map of the Netherlands.

## Data source for corona data files

sewage data in json format and number of municipalities cumulative in json format.

both can be found on:

https://data.rivm.nl/covid-19/

## Data source for the geomapping

From the pdok website the provinces outline were downloaded in shp format.

The data for this can be found on:

https://www.pdok.nl/

## With this data the following questions have be answered:

**- Do the corona cases have a similar curvation throughout the time per province**

This will also be tested by visualization and doing a statistical test. First a heatmap will be made showing the cases per province. Next a statistical test will be performed based on the normality of the curvation of the data this will either be ANOVA or Kruskal-Wallis H-test.

**- Does the sewage data correspond with the data from the corona cases?**

This will be tested by visualization of both the sewage data and the amount of corona cases in the provinces Groningen and Friesland.

- Additionally a map of the Netherlands the amount of measured corona cases from each province wil be visualized.

## Set up of the experiment

To answer the research questions the data will have to be preprocessed first by cleaning up the data and removing unnecessary data. Furthermore the sewage data does not contain the provinces but the regions so these have to be changed into the correct provinces.

To compare the datasets with each other the datasets were both visualized in a heatmap per province to see if they follow the same trend in the number of corona cases.

Next the corona cases were checked on distribution. and a statistical test was performed.

To check out Groningen and Friesland, these provinces were isolated from the rest of the dataset. Next a sum was made based on the date for each province and the data was normalized to get a better representation when comparing them with each other. Next two simple plots were made for the provinces Groningen and Friesland. To get a better view a correlation was done on te total corona cases and RNA flow in sewage. Lastly the data was checked on distribution and a statistical test was done.

After answering the research questions three maps were made of the corona cases per province. The three consist out of, the total amount of corona cases, hospitalization due corona, and deceased due corona.

## Conclusion and Discussion

- Firstly looking at the first research question, do the corona cases have a similar curvation throughout time per province.

It seems to not have a similar curvation, when looking at the heatmap, the statistical test and the histograms, it looks like although they show some similarities it appears to still be very different per province.

- The second research question, does the sewage data correspond with amount corona cases in Groningen and Friesland?

I was very excited about this research question. But was very disappointed to not find any good data in the sewage data. The amount of measurements taken per province and the times they are taken are very different. Also some measurements seem to be extremely high at one moment and then very low at the next. This seems very unlikely when looking at the total amount of corona cases. So to answer this question, no the sewage data does not correspond with the amount of corona cases in Groningen and Friesland. Although this experiment could have been wider, looking at all the provinces. Looking at the data this does not seem to change anything.