Practical: GIS for Health

- Duration: 3 hours
- Software needed: QGIS 3.10.4

Introduction:

The goal of the practicum is to get familiar with spatial data, geo-information systems, mapping and spatial analyses. After you have finished the exercises you should have a general idea about the potential use of geo-information in the domains of health. More specific you will:

- Know how to find interesting spatial data
- Know how to add them to a Geographic Information System
- Know how to combine spatial data with additional sources of data
- Know how to produce a (thematic) map
- Know how to do a based spatial analysis

To prepare:

- Follow the introduction lecture
- Install QGIS 3.10 (64 bit) (either through the available software app of Wageningen University or https://www.qgis.org/en/site/forusers/download.html)
- Have a look at the first 10 min. of https://www.youtube.com/watch?v=kCnNWyl9qSE

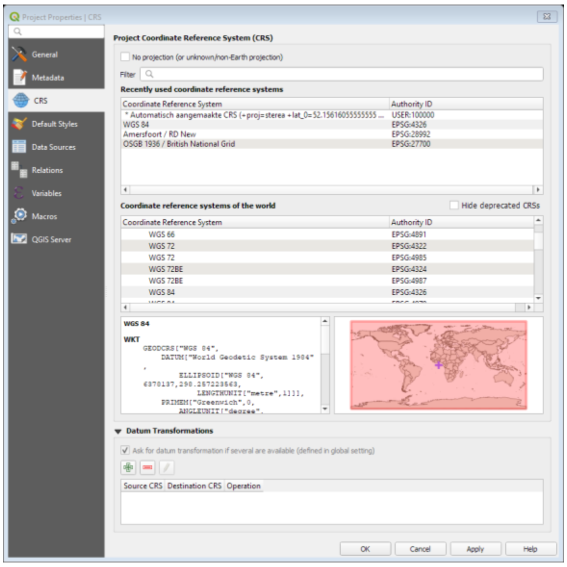
Exercise 1: Create a basic thematic map

In this first exercise you will create a map with the distribution of registered COVID-9 infections for the Dutch municipalities. This is called a choropleth map by cartographers, meaning that areas (mostly administrative area such a municipalities, neighborhoods, zip-code areas etc.) are colored/shaded proportional to a numerical variable.

The first step (in almost all GIS analyses) is to make sure you use a proper coordinate system [1]in your geographic information system (GIS). The default setting of QGIS is WGS84 (Google what this means). However, the official coordinate system used in the Netherlands the “Amersfoort/RD-new” or “EPSG28992”. Most, if not all, official dataset in the Netherlands a defined using this coordinate system. It’s good to realize that each country in the world uses it’s own system. To make sure your project has the right coordinate system take the following steps:

1. Start QGIS 3.10
2. Start a new project by choosing “new” from the “project” menu (or type crtl-n)
3. Choose “properties” form the project menu and the following interface will show:

[1] It’s not necessary for this exercise to understand exactly how coordinate systems work but if you want to know see: https://en.wikipedia.org/wiki/Geographic_coordinate_system or look at



Project Coordinate Reference System (CRS)

□ No projection (or unknown/non-Earth projection)

Filter

Recently used coordinate reference systems

Coordinate Reference System	Authority ID
* Automatisch aangemaakte CRS (+proj=sterea +lat_0=52.15616055555555 ...	USER:100000
WGS 84	EPSG:4326
Amersfoort / RD New	EPSG:28992
OSGB 1936 / British National Grid	EPSG:27700

Coordinate reference systems of the world

□ Hide deprecated CRSs

Coordinate Reference System	Authority ID
WGS 66	EPSG:4891
WGS 72	EPSG:4322
WGS 72	EPSG:4985
WGS 72BE	EPSG:4324
WGS 72BE	EPSG:4987
WGS 84	EPSG:4326

WGS 84

WKT

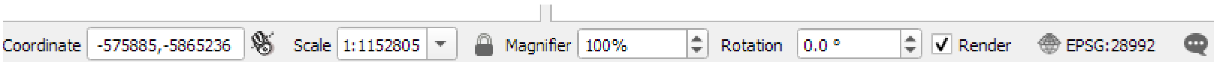
```
    GEODCRS["WGS 84",
        DATUM["World Geodetic System 1984"
    ,
            ELLIPSOID["WGS 84",
    6378137,298.257223563,
                LENGTHUNIT["metre",1]]],
        PRIMEM["Greenwich",0,
            ANGLEUNIT["degree",
```

Datum Transformations

☑ Ask for datum transformation if several are available (defined in global setting)

Source CRS	Destination CRS	Operation

4. Find the Amersfoort / RD new coordinate reference system (tip: use the filter box). Make sure you choose ESPG28992 since there are different versions.
5. After selecting, click ok and your project is using the right coordinate system. You can check this by looking at the right corner of the window where it should say EPSG:28992



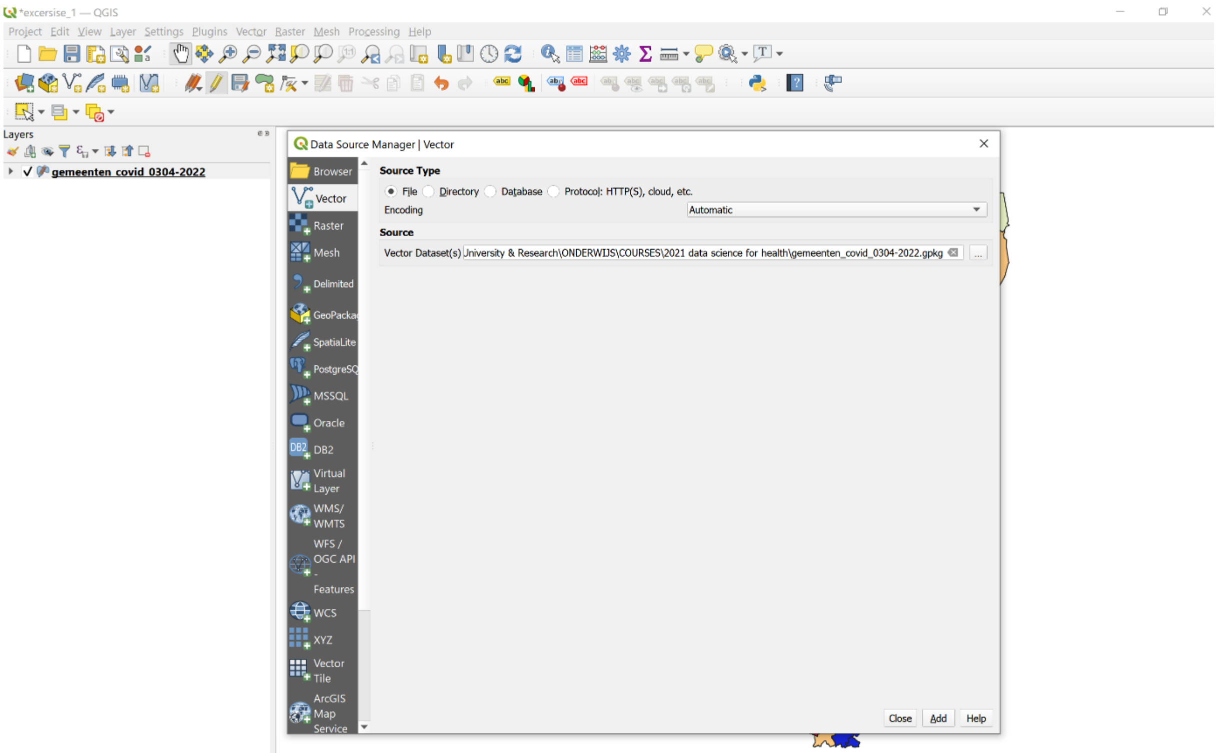
Coordinate -575885,-5865236 Scale 1:1152805 Magnifier 100% Rotation 0.0 ° ☑ Render EPSG:28992

Now we have set our GIS to the proper coordinate system we can add data. There a many ways to find data. For this assignment we will use the site **data.overheid.nl** which contains a comprehensive overview of the data made available by the Dutch governmental organizations. Searching for "Covid-19" yields the dataset of the RIVM with cumulative number of infections in the Dutch municipalities (called: "Covid-19 cumulatieve aantallen per gemeente").

Todo: Go to this website and locate the dataset described above. Next check the description of the dataset. Find out what information is present in the dataset.

The website offers different way to obtain the data. You can download the data as comma delimited file (csv) or JSON. Since it is not a spatial data set (why not?). We have converted this data for you to a spatial dataset you can download from BrightSpace. It's called "gemeenten_covid_03_04-2022.gpkg".

1. Download the gpkg file from Brightspace. Next in QGIS go to "Layer" --> "Add vector layer" and open "gemeenten_covid_0304-2022.gpkg".



Source Type

● File ○ Directory ○ Database ○ Protocol: HTTP(S), cloud, etc.

Encoding: Automatic

Source

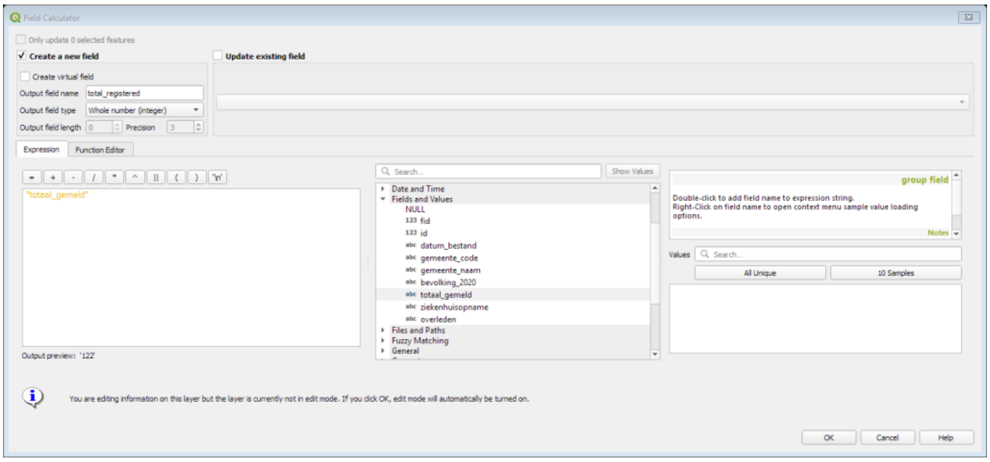
Vector Dataset(s): University & Research\ONDERWIJS\COURSES\2021 data science for health\gemeenten_covid_0304-2022.gpkg

2. The result will be a map with the Dutch municipalities in a uniform color. To inspect the data you might want to **right-click** on the dataset name and select "Open attribute table"

Question: open the attribute table and check what information in inside. How does it compare to the information on the data.overheid.nl site?

For sure this is not the map you like, as it does not show interesting information yet. To create a map that indeed shows information about Covid-19 infections we need to assign colors to each municipality that relates to the counts of infected people. As you can see in the attribute table there are 2 columns (fields) that contain this information (which ones?). As an example we will create a map showing the information in the field "covid03_04_2022_total_reported" that indicate the registered Covid-19 infections. A problem with this field is that the information in the table is in a textual format. You can check this by right-clicking the dataset, choose "Properties" and next choose "Fields" tab. You can see the information in the "covid03_04_2022_total_reported" field is of type String (which is computer jargon for textual data). Textual information cannot but converted to classes directly; it first need to be converted to a numeric format.

1. To convert the textual data to a numeric format open the attribute table of the dataset and open the "Field calculator" by pressing
2. In the field calculator choose "Create a new field" and enter a name for the new field (for example "infected") and choose as output field type " Whole number (integer)" . Next type in the expression field "covid03-04-2022_total_reported" (use quotes) (you also can pick it from "Fields and Values"). This make that the field "covid03-04-2022_total_reported" is copied to the new field as a numeric value.
3. Save the changes in the attribute table by clicking and close the table.



Field Calculator

□ Only update 0 selected features

☑ Create a new field

□ Create virtual field

Output field name: total_registered

Output field type: Whole number (integer)

Output field length: 0 Precision: 3

□ Update existing field

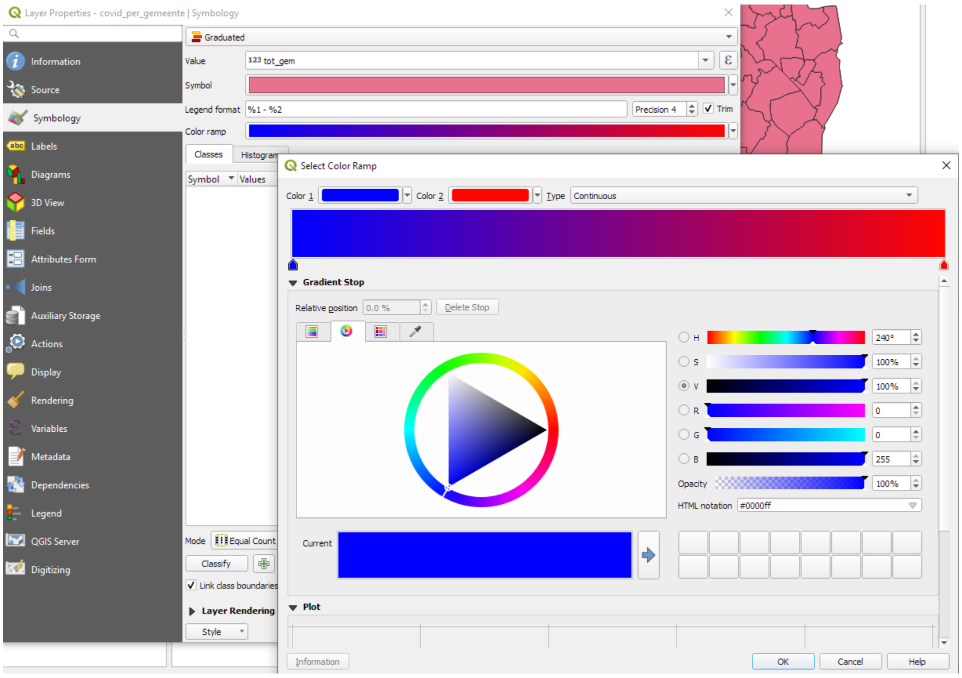
Expression: "totaal_gemeld"

Output preview: '122'

Double-click to add field name to expression string.
Right-Click on field name to open context menu sample value loading options.

You are editing information on this layer but the layer is currently not in edit mode. If you click OK, edit mode will automatically be turned on.

4. To create a classified map showing the number of registered Covid-19 infections double-click the layer and the “Symbology editor will open (alternatively right-click the layer and choose “properties” and the Symbology tab)
5. Change the dropdown that says “Single symbol” and change it to “Graduated”
6. Choose from the value dropdown the just created field with the numeric values of the registered infections (total_registered)
7. Under Color Ramp choose the option “Choose New color ramp” of the type Gradient and define a color ramp (for example from blue to red). Click “OK”



Layer Properties - covid_per_gemeente | Symbology

Graduated

Value: 123 tot_gem

Legend format: %1 - %2 Precision 4 ☑ Trim

Mode: Equal Count

☑ Link class boundaries

Select Color Ramp

Color 1 Color 2 Type: Continuous

Relative position: 0.0 %

H 240° S 100% V 100% R 0 G 0 B 255 Opacity 100%

HTML notation: #0000ff

8. Choose from “Mode” one of the classification methods. Define in “classes” the number of classes you would like to define. Depending on the classification methods you can manually change the class boundaries if you like. Experiment with the number of classes en classification methods
9. After finishing click “apply” or “ok” and inspect the result

The map created this way does not yield a completely representative figure of the outbreak since it does not take into account the number of inhabitant of a municipality. To create a better figure we can normalize the data based on the inhabitants of a municipality:

Question: What would you think is a proper way to normalize the data such that municipalities can be compared?

1. Open the attribute table and Field Calculator.
2. Add a new numeric field (decimal number) and populate it by entering the following expression in the expression field: **(to_int("totaal_gemeld")/to_int("bevolking_2020"))*100**

Question: what does the function to_int("layer name") mean?

3. Save the table and close.
4. Update the map by creating a new classification based on these relative numbers. First delete the previous classification by hitting the "Delete All" button in the Symbology editor form. Next choose the field (columns) containing the normalized values.

Tips:

- To quickly check the descriptive statistics of a field goto "View"--> "Panels" and check "Statistics Panel"
- You can label the municipalities by right-clicking the layer, choose properties and the tab "Labels". Next choose single label a choose the gemeente_naam field as Value. Next adjust the font and font size etc.

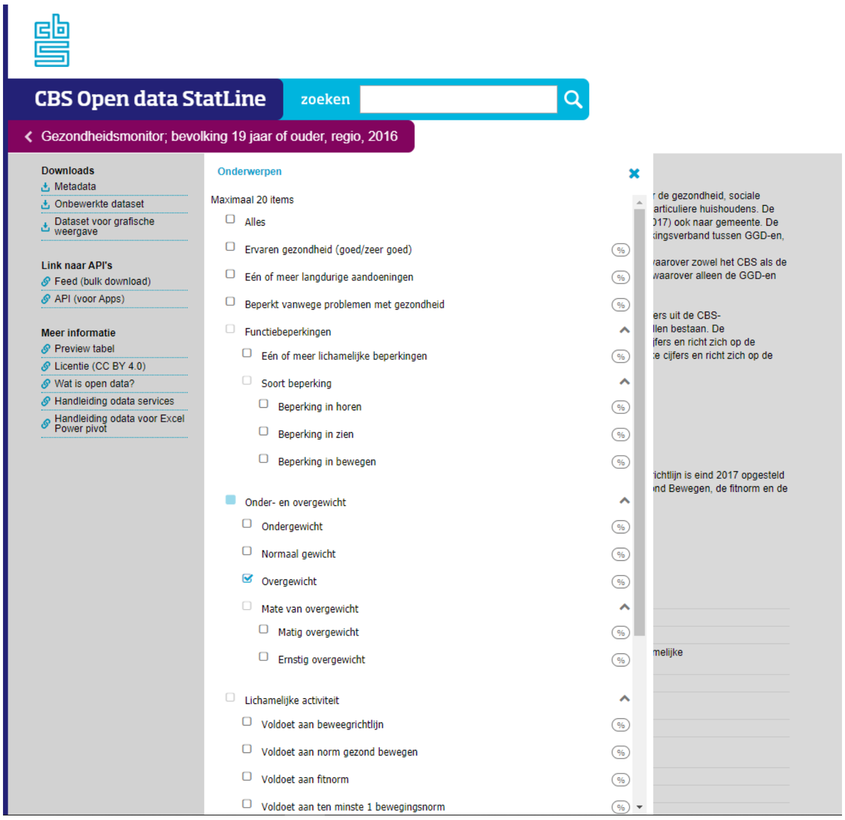
-If time permits you might also want to create a map with people admitted to the hospital

Exercise 2: Combine information

In this exercise you will practice with combining various datasets (spatial and non-spatial). As a case study we will create a map showing the number of obese persons in the Netherlands. We therefore will combine the dataset of the municipalities (of the previous exercies) with a dataset of the central bureau of statistics (CBS)

Download statistical data about obesity. This information is present at the Dutch Central Bureau of Statistic (CBS).

1. Got to https://opendata.cbs.nl/statline/portal.html?_la=nl&_catalog=CBS (which provide opendata access to all CBS tables) and select "Nederland Regionaal" --> "Gezondheid, leefstijl,zorggebruik" --> "Gezondheidsmonitor; regio, 2020"
2. Click "onbewerkte dataset" and start assembling your dataset. We limit our self and only choose in "Onderwerpen": "Overgewicht" and in "Leeftijd: 18-65" and "Regio's" only "gemeenten" .



CBS Open data StatLine

zoeken

Gezondheidsmonitor; bevolking 19 jaar of ouder, regio, 2016

Downloads
- Metadata
- Onbewerkte dataset
- Dataset voor grafische weergave

Link naar API's
- Feed (bulk download)
- API (voor Apps)

Meer informatie
- Preview tabel
- Licentie (CC BY 4.0)
- Wat is open data?
- Handleiding odata services
- Handleiding odata voor Excel Power pivot

Onderwerpen

Maximaal 20 items

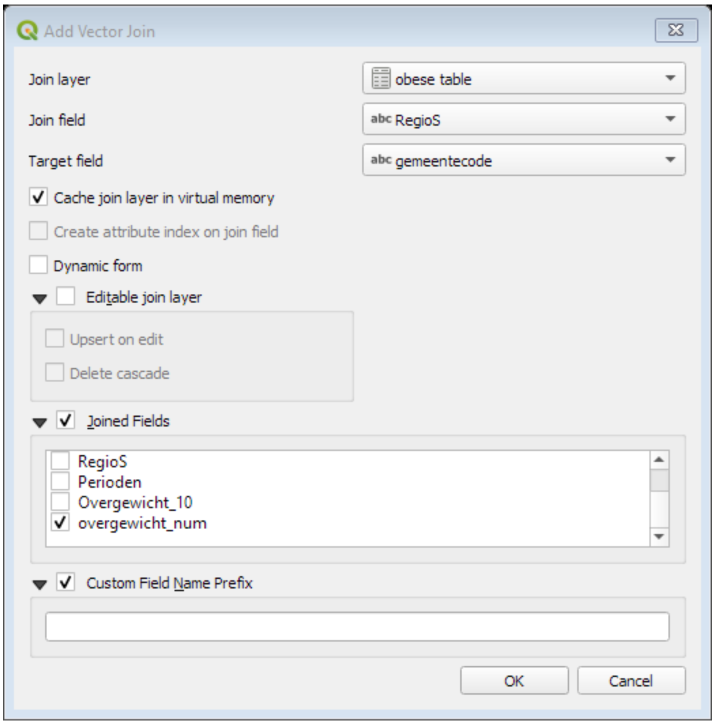
- □ Alles
- □ Ervaren gezondheid (goed/zeer goed)
- □ Eén of meer langdurige aandoeningen
- □ Beperkt vanwege problemen met gezondheid
- □ Functiebeperkingen
  - □ Eén of meer lichamelijke beperkingen
  - □ Soort beperking
    - □ Beperking in horen
    - □ Beperking in zien
    - □ Beperking in bewegen
- ■ Onder- en overgewicht
  - □ Ondergewicht
  - □ Normaal gewicht
  - ☑ Overgewicht
  - □ Mate van overgewicht
    - □ Matig overgewicht
    - □ Ernstig overgewicht
- □ Lichamelijke activiteit
  - □ Voldoet aan beweegrichtlijn
  - □ Voldoet aan norm gezond bewegen
  - □ Voldoet aan fitnorm
  - □ Voldoet aan ten minste 1 bewegingsnorm

3. Next click “Download CSV”.
4. Back in QGIS choose “Layer--> “Add Layer” --> “Add Delimited Text layer”.
5. Choose the file you just downloaded and change the layer name if you like.
6. Change to “Custom delimiters” pick “Semicolon” as delimiter. Leave the rest as default.
7. Click “Add” and “Close”.
8. Open the attribute table of the just added table and check what’s inside. If you hover over the fieldnames you will see that the field containing the “overgewicht” values is in a textual format. **Add an extra column that contains the values in a numeric format** following the same procedure as in assignment 1 (take a decimal number as format).

To connect (join) two datasets (in this case the spatial layer with the municipalities and the table with obese percentages) a common field is required. For our case this is the municipality code which is present in both datasets (in the obese table it’s called “**RegioS**” and in the). Each municipality in the Netherlands has a unique code which can be used to connect to RegioS field.

Question: Why is it important that the code is unique in the municipality dataset? What could happen if it is not?

9. Right-click the layer with the municipalities and choose “Properties” and the “Joins” tab and the “+” sign.
10. Choose as “Join layer” the table with obese information, as “Join field” RegioS and a “Target Field” “gemeente code”. Choose as “Joined Fields” the filed with the numeric representations of the overweighed percentage of people (Remove the “Custom Field Name Prefix”)



Add Vector Join

Join layer: obese table

Join field: abc RegioS

Target field: abc gemeentecode

☑ Cache join layer in virtual memory

□ Create attribute index on join field

□ Dynamic form

□ Editable join layer

□ Upsert on edit

□ Delete cascade

☑ Joined Fields

□ RegioS

□ Perioden

□ Overgewicht_10

☑ overgewicht_num

☑ Custom Field Name Prefix

11. Click “OK” and inspect the attribute table of the municipalities layer

Question: You see for some municipalities a value of *NULL* in the column containing the overweight percentages. What does this mean and how can this happen?

12. Now you can create a choropleth map simmilar like you did during exercise 1 using a classification (try Natural Breaks (Jenks)) and graduated color scheme.

Exercise 3: A basic analyses

This exercise will show you how perform a basic spatial analysis using GIS software. As a case study we will reconstruct the analyses of the Cholera crises in London in 1854 by John Snow, generally seen as one of the first applications of spatial data in an epidemiological study. (see: https://en.wikipedia.org/wiki/John_Snow).

1. Download the Geopackage file (exercise_3.gpkg) with data on the cholera outbreak in 1854 from Brightspace.
2. Start a new project in QGIS and set the coordinate system to “British National Grid (EPSG 27700)
3. Open the original Snow Map (SnowMap.tif). These is a background image (raster). Use the Data Source Manager by clicking (ctrl L) and choose “Raster”. Next open the Geopackage file of exercise 3 (Exercise_3.gpkg).
4. Open the spatial data sets with location of the pumps (pumps) and the locations and counting of people who died from cholera (Cholera_deaths). These files are vector files (points) and also stored in the Geopackage. Use “Vector” tab in the Data Source Manager.
5. You might want to change the colors and/or symbols of the vector files using the symbology editor by double clicking on the layer name (or open the styling panel by clicking or press F7). Change for example the symbol of the pump to a blue triangle.
6. Open the attribute tables to the “pump” and “cholera_deaths” layers and explore what information is present.

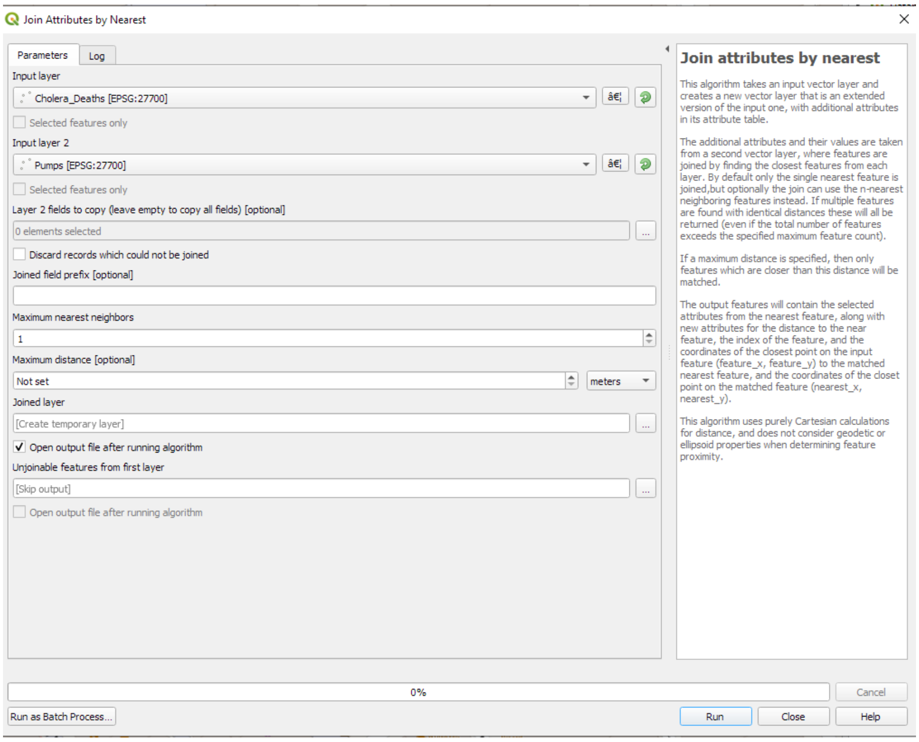
During his analyses John Snow discovered that most Cholera victims were locate near a specific pump in Broad Street. He analyzed it by creating a map with bars indicating the number of deaths at each location (check this on the original map). Each bar represented one cholera victim.

Using a GIS analyses we can 'reproduce' his findings. The follow procedure is proposed: 1) Associate the locations of the registered cholera victims to the closest pump and 2) next calculate the number of victims associated with that pump. To do so:

7. If not opened yet open the "Processing Toolbox Panel" by choosing "Processing" --> "Toolbox" from the menu.
8. In the toolbox go to to "Vector general" and choose "join attributes by nearest"

Question: What is this command doing?
See:https://docs.qgis.org/3.10/en/docs/user_manual/processing_algs/qgis/vectorgeneral.html#join-attributes-by-nearest

9. Take as the first Input layer "Cholera_deaths" and as Input layer 2 "Pumps" and accept the rest as default



Input layer
Cholera_Deaths [EPSG:27700]
□ Selected features only
Input layer 2
Pumps [EPSG:27700]
□ Selected features only
Layer 2 fields to copy (leave empty to copy all fields) [optional]
0 elements selected
□ Discard records which could not be joined
Joined field prefix [optional]
Maximum nearest neighbors
1
Maximum distance [optional]
Not set meters
Joined layer
[Create temporary layer]
☑ Open output file after running algorithm
Unjoinable features from first layer
[Skip output]
□ Open output file after running algorithm

Join attributes by nearest

This algorithm takes an input vector layer and creates a new vector layer that is an extended version of the input one, with additional attributes in its attribute table.

The additional attributes and their values are taken from a second vector layer, where features are joined by finding the closest features from each layer. By default only the single nearest feature is joined,but optionally the join can use the n-nearest neighboring features instead. If multiple features are found with identical distances these will all be returned (even if the total number of features exceeds the specified maximum feature count).

If a maximum distance is specified, then only features which are closer than this distance will be matched.

The output features will contain the selected attributes from the nearest feature, along with new attributes for the distance to the near feature, the index of the feature, and the coordinates of the closest point on the input feature (feature_x, feature_y) to the matched nearest feature, and the coordinates of the closet point on the matched feature (nearest_x, nearest_y).

This algorithm uses purely Cartesian calculations for distance, and does not consider geodetic or ellipsoid properties when determining feature proximity.

10. Open the attribute table of the newly create (temporary) layer (default name is "Joined layer", you might want to rename it) check if you understand what is inside now

Question: what does Id and Id_2, and the distance refer to?

11. Change the symbology of the newly created layer by assigning a color based on the Id_2 field. (Open the Symbology editor or panel and choose for "Categorized", choose the I_2 field and hit the "Classify" button).

Questions: What does it mean what you see now? Why are not all pumps (Id_2) present?

12. Go to "vector analyses" in the processing toolbox
13. Choose Statistics by categories and choose the "Joined layer" as the input vector layer

14. Use the "Count" field as the field to calculate the statistics on and take the Id_2 field as category field (why?) and hit the run button.
15. Open the newly create table (not a spatial dataset) and inspect it.

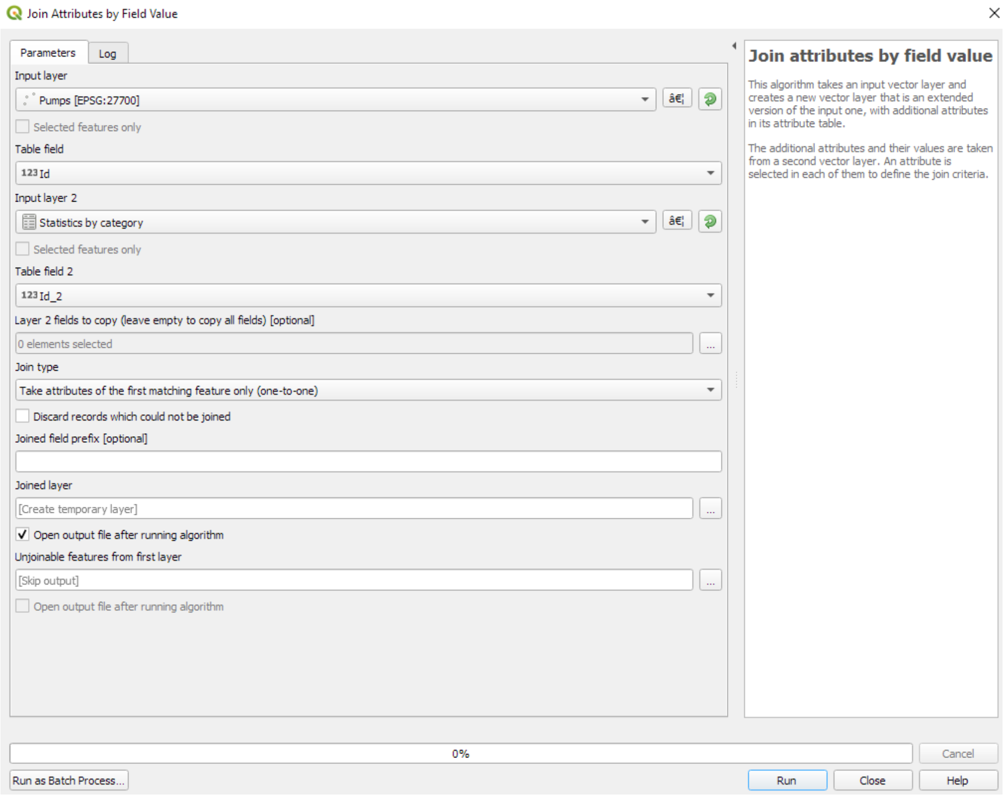
Questions: what is the id of the pump near the locations with the most cholera outbreak? And in which street is it located?

Extra:

Create a map with proportional symbols:

It is possible to create a map with symbols of different sizes indicating the number of deaths associated with a pump by:

1. Choose "join attribute by field value" from "Vector General" in the Processing toolbox
2. Take as input layer the pumps layer and the Id field as table field and as input layer 2 the statistics by category table and the Id_2 field and hit the "Run" button.



Join Attributes by Field Value

Parameters | Log

Input layer: Pumps [EPSG:27700]

□ Selected features only

Table field: 123 Id

Input layer 2: Statistics by category

□ Selected features only

Table field 2: 123 Id_2

Layer 2 fields to copy (leave empty to copy all fields) [optional]: 0 elements selected

Join type: Take attributes of the first matching feature only (one-to-one)

□ Discard records which could not be joined

Joined field prefix [optional]:

Joined layer: [Create temporary layer]

☑ Open output file after running algorithm

Unjoinable features from first layer: [Skip output]

□ Open output file after running algorithm

Join attributes by field value

This algorithm takes an input vector layer and creates a new vector layer that is an extended version of the input one, with additional attributes in its attribute table.

The additional attributes and their values are taken from a second vector layer. An attribute is selected in each of them to define the join criteria.

3. A new layer is generated called "Joined Layer". Open the attribute table and inspect what is inside.

Question: do you see that some wells have NULL values? Which are they and why?

4. Open the Symbology Editor by double clicking the joined layer
5. Take as value field the "sum"
6. Select "Graduated" as symbol type and change "method" to "Size"
7. Click the "Classify" button to see how it looks. Play with the classification method, colour and size of the symbols if you like

Create a heatmap

To create an impression of the occurrences of a certain spatially dispersed phenomenon a “heatmap” is often used. A common way to create a heatmap is to apply a so called “Kernel Density Estimation” algorithm which is basically a spatial representation of a probability density function (see for a detailed explanation: https://pro.arcgis.com/en/pro-app/tool-reference/spatial-analyst/how-kernel-density-works.htm). Creating a heatmap in QGIS is simple:

1. In the “Processing Toolbox” go to “Interpolation” and pick “Heatmap (Kernel Density Estimation)”
2. Choose as point layer the Cholera_deaths and take a radius of 25 or 50 m. and hit “OK”
3. To enhance the heatmap open the “Symbology” editor, and choose “ Singleband pseudocolor” as render type and pick an colorramp. Experiment with the rendering options.