Exercise

Analysis of John Snow's 1854 cholera data

Julian Flowers

2024-04-20

Contents

Introduction 1
Get started 1
Get the data 2
Distribution 3
Map the data 3
Add analysis 4

Introduction

In this exercise we will use data from the 1854 cholera outbreak in London which was investigated by a local doctor, Dr John Snow, who is widely regarded as a founding father of modern epidemiology. His observational work and analysis helped stop the outbreak but also identified the water borne nature of cholera outbreaks.

In this exercise we will be using R and Tableau to analysed and visualise the data. The source data is available here.

Get started

First we need to load the R packages for analysis and GIS

```
needs(sf, stars, tidyverse, mapview)
```

Now we load the data. The spatial points (vectors) for cholera deaths and pump locations are stored in a specialised spatial data format known as a shapefile (.shp), and the background maps as rasters (.tif). In R we can read the points with the `read_sf` function in the `sf` package, and the rasters with `read_stars` from the `stars` package.

These data are the:

- OSMap Raster Modern OS map of the area of the outbreak (from OS Open Data - contains Ordnance Survey data © Crown copyright and database right 2013). Ordnance Survey is the main spatial data provider in the UK.

- OSMap_Greyscale Raster Same as above, but in greyscale for easier visualisation (altered by conversion to greyscale, from OS Open Data - contains Ordnance Survey data © Crown copyright and database right 2013)
- SnowMap Raster Snow's original map, georeferenced and warped so that it accurately overlays the OS map
- CholeraDeaths Vector Points for each location of one or more deaths. Attribute value gives number of deaths at that location
- Pumps Vector Points for each location of a pump

Get the data

The code segment below shows how data is loaded.

```
deaths <- read_sf(paste0(path, "/Cholera_Deaths.shp"))

deaths_coords <- sf::st_coordinates(deaths) |> data.frame()

pumps <- read_sf(paste0(path, "/Pumps.shp")) |>
    mutate(id1 = row_number())

osm1 <- stars::read_stars(paste0(path, "/SnowMap.tif"))
osm2 <- stars::read_stars(paste0(path, "/OSMAp_Grayscale.tif"))
```

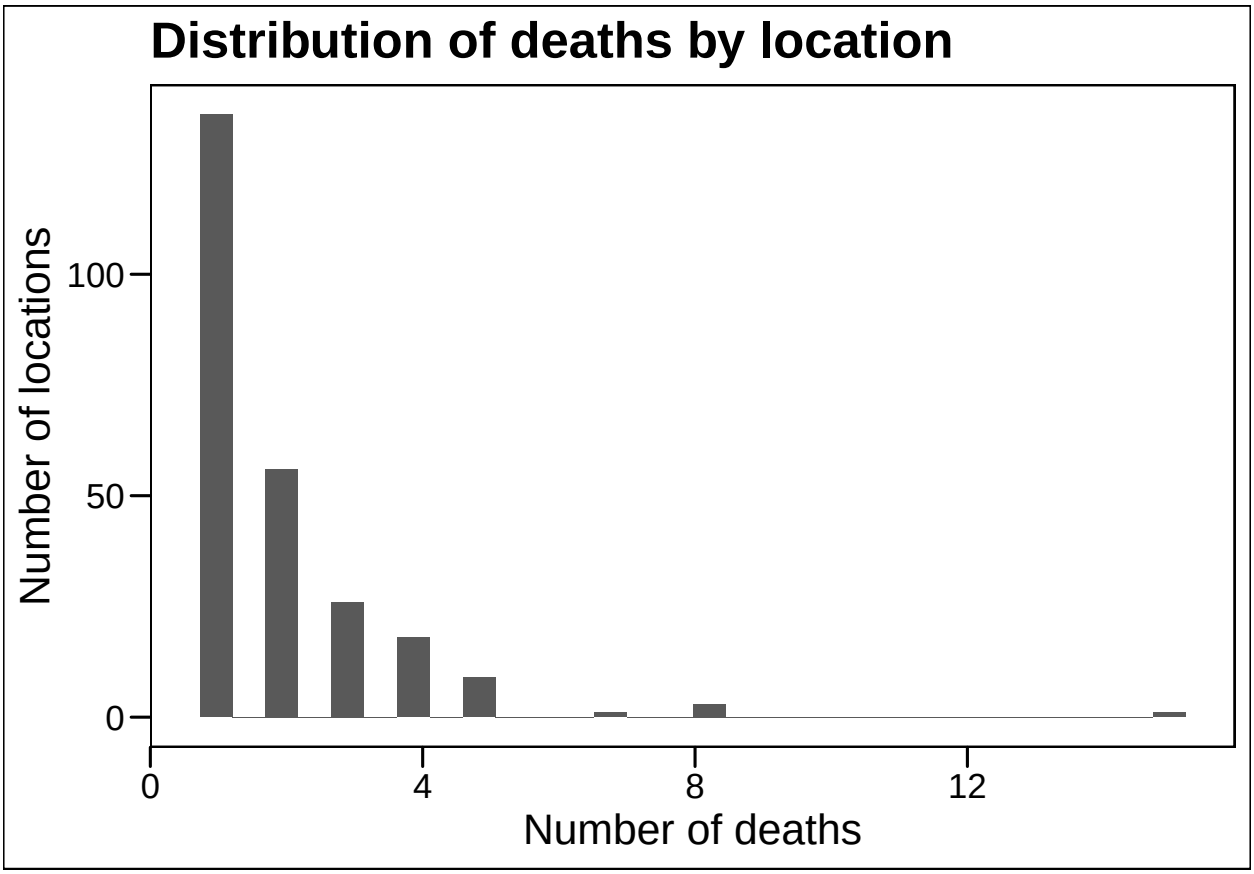
```
## Simple feature collection with 250 features and 2 fields
## Geometry type: POINT
## Dimension:     XY
## Bounding box:  xmin: 529160.3 ymin: 180857.9 xmax: 529655.9 ymax: 181306.2
## Projected CRS: OSGB36 / British National Grid
## # A tibble: 250 x 3
##       Id Count            geometry
##    <int> <int>         <POINT [m]>
##  1     0     3 (529308.7 181031.4)
##  2     0     2 (529312.2 181025.2)
##  3     0     1 (529314.4 181020.3)
##  4     0     1 (529317.4 181014.3)
##  5     0     4 (529320.7 181007.9)
##  6     0     2   (529336.7 181006)
##  7     0     2 (529290.1 181024.4)
##  8     0     2   (529301 181021.2)
##  9     0     3   (529285 181020.2)
## 10     0     2 (529288.4 181031.8)
## # i 240 more rows
```

If we review the `deaths` data we can see it consists a set of metadata which tells us the type of geometry (point), the dimensions (X and Y) the limits of the area and the coordinate reference system (CRS). In this case the the locations are measured in meters but they can be converted to degrees.

The first 10 records are shown - we can see there is an Id, a Count - the number of deaths at each location - and a geometry field which contains the point location. In all there are 250 locations where deaths occurred.

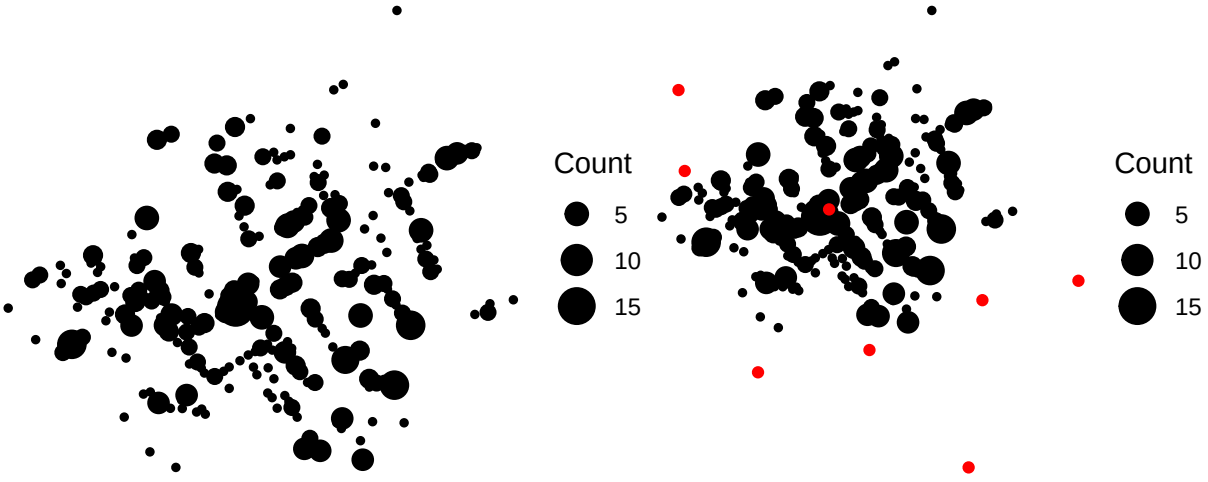
Distribution

We can look at the distribution of death counts by plotting a histogram.

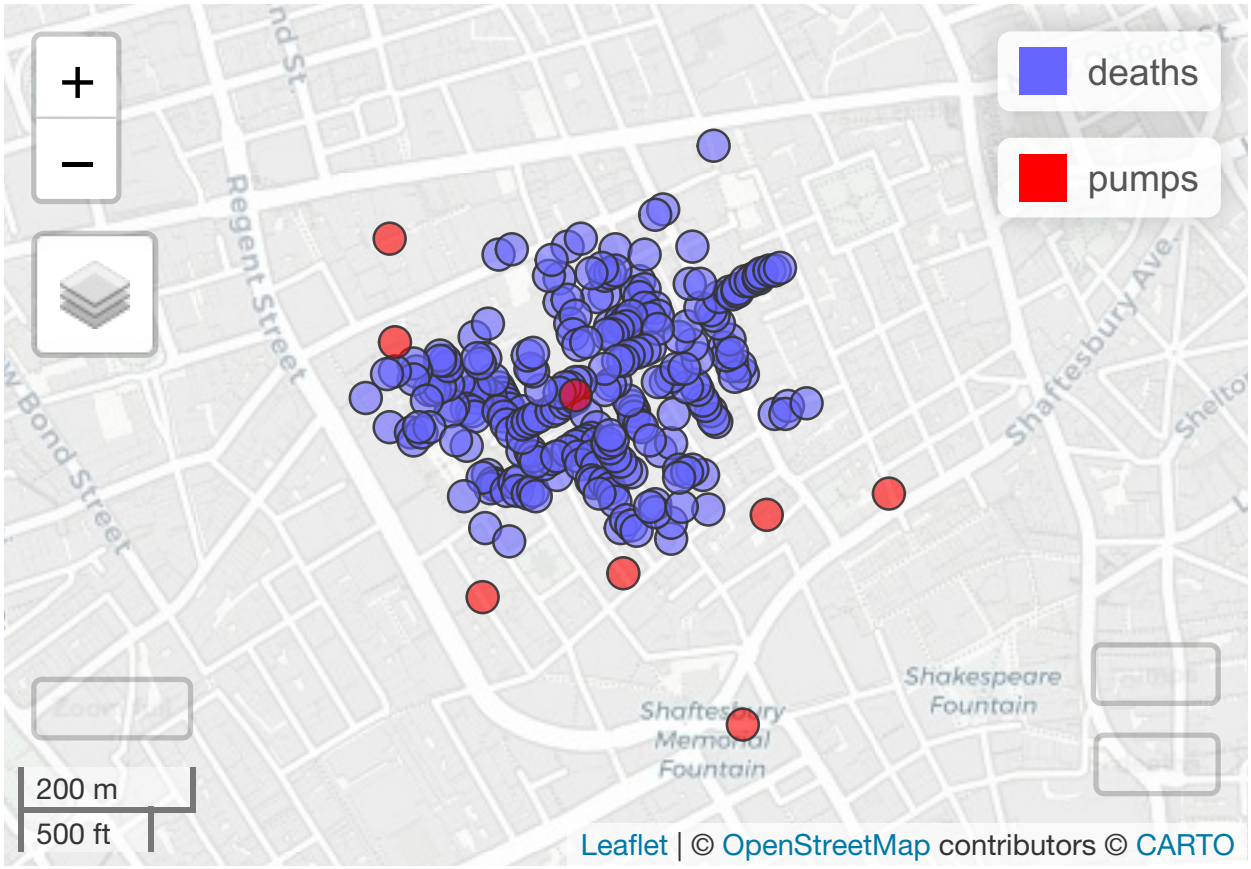


This shows that at the majority of locations there was only 1 death, but there were many sites with multiple deaths. If we plot the locations on a map

Map the data



We can create an interactive map of the data using `mapview`...



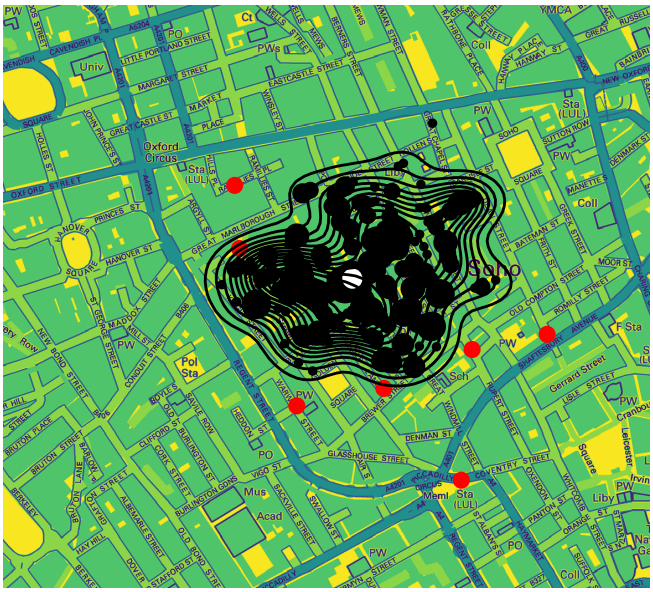
Add analysis

Finally overlaying the point data with kernel density estimate (kde) can provide a succinct visualisation of the spatial density of point data.

```
deaths |>
    ggplot() +
    geom_stars(data = osm2, show.legend = FALSE) +
    geom_sf(aes(size = Count), show.legend = FALSE) +
    geom_sf(data = pumps, colour = "red", aes(size = 2), show.legend = FALSE) +
    geom_sf(data = pumps |> filter(id1 == 1), colour = "white", aes(size = 3), show.legend = FALSE) +
    geom_density2d(aes(deaths_coords$X, deaths_coords$Y), colour = "black", show.legend = FALSE) +
    theme_void() +
    theme(plot.title.position = "plot") +
    scale_fill_viridis_d(option = "viridis") +
    labs(title = "Location of cholera deaths in 1854 outbreak in relation to water pumps",
         subtitle = "Overlaid with a 2D kernel density estimate which shows\nthe Broad(wick), pump at t
         caption = "Death: Black dot\nEpicentre pump: White dot\nRed dot: Pump") +
    theme(title = element_text(size = 12))
```

Location of cholera deaths in 1854 outbreak in relation to w

Overlaid with a 2D kernel density estimate which shows
the Broad(wick), pump at the epicentre of the outbreak



Death: Black dot
Epicentre pump: White dot
Red dot: Pump

This shows how one pump (at what is now called Broadwick Street) is at the epicentre of the outbreak.