Because the health data is identified by clinical commission group (CCG) and the pollution data by borough, we had to streamline this to create one common variable to connect the data over time. To do this, we had to first identify what pollution monitoring sites in each borough were open and had available data. Next, we had to place each monitoring site in the locations in question with their respective CCG. Since for our research we made a distinction between an inner and outer radius from Heathrow, we had to decide where to place the monitoring sites that were between CCGs. We have done such until now. The next steps are to download the data.

Because there is a row limit of 11,000 to download the pollution data, we have to download the data, which is available per monitoring site by the hour, for every monitoring site required. We then have to concatenate the data to create one file per monitoring site. We will do this by creating a folder with the inputs, concatenate using the skills we have learned in Jupyter Notebook, and creating a separate folder with the concatenated output files. Then, we have to clean the data. This means deleting the inessential header rows that are not data. Next, we check for null values and determine what to do with them. By this time, each monitoring site will align with a CCG, so we will add a new column using the CCG to describe the location. We will use this column instead of the borough name to connect the pollution data with the health data.