This document outlines the precise meaning, and source, of all columns in the Cases\_Data.csv

Each row of the dataframe refers to a specific LTLA, at a specific week. All other columns are related contextual information for that LTLA-week.

It’s worth nothing that for nearly all of these covariables, there were a few edge cases where some LTLAs weren’t captured exactly due to boundary changes. In these instances I would often taken average of the neighbour, or do some other type of aggregating or averaging depending on the context.

Code that put this “Cases\_Data.csv” together can be provided, and was built as part of this project: https://doi.org/10.1101/2024.04.20.24306121

**areaCode** – There is a unique code assigned to every imaginable geography in England. You can learn more about the various types of these codes here: <https://geoportal.statistics.gov.uk/datasets/ce24654e47e94906ae749cd9741ec318/about>

This is the unique areaCode assigned to this row’s LTLA. Code is a more reliable “link” than name, because it avoids unfortunate differences in name or spelling.

**Week** – This is a number from 2-131 referring to which “week” of the pandemic we’re referring to. Week 1 here refers to the w/b May 3rd 2020. This is my own personal convention, and may not link up with other people’s definition of “epidemic week”. This was picked as it was considered “the first week of reliable testing” for my own research projects. You may choose to ignore this convention!

**areaName** – The name of the LTLA in question.

**Population** – Population of LTLA according to mid-2020 estimates. See “ukpopestimatesmid2020on2021geography.xlsx”

**Pop\_per\_km2** – The population per square kilometre of the LTLA as also taken from mid-2020 estimates. See MYE5 of “ukpopestimatesmid2020on2021geography.xlsx”.

**Median\_age** – The median age of the LTLA population, as also taken from mid-2020 estimates. See MYE6 of “ukpopestimatesmid2020on2021geography.xlsx”.

**INDEX** – This is a unique index number, 1:356, I assigned to each LTLA to make things easier when I was doing some matrix calculations, you could probably ignore.

**Week\_Cases** – This is the number of reported COVID-19 cases from within this LTLA for this week. Aggregated from the daily data from the UK dashboard. This dashboard seems to be down now, but the api link was: (<https://api.coronavirus.data.gov.uk/v2/data?areaType=ltla&metric=newCasesBySpecimenDate&format=csv>)

I have included the latest saved download of the API as ltla\_cases\_2023\_06\_22.csv

Note that this is by *specimen date*, i.e. the date the test was performed, not the date that the result was reported.

**date\_begin** – This is the first calendar date of the week this row refers to.

**Previous\_week\_cases / next\_week\_cases** – This is just the previous and next week’s aggregate cases. This was just included to make some calculations easier for me.

**First\_Episodes\_Total** – The rolling cumulative sum of the number of reported “first episodes” up to the first date of the week as provided from the dashboard, via PCR or LFD, pillar 1 or pillar 2. Taken from dashboard, source file ltla\_first\_episodes\_2023\_06\_22.csv

**cumVaccPercentage\_FirstDose** – The percentage of this LTLA’s population that has had their first dose of vaccine by the start of this week. This is taken directly from the dashboard (ltla\_cum\_1dose\_vaccine\_percentage\_10\_11\_2022.csv)

**Linelist\_P2\_PCR\_Week\_Cases** – These were a different measures of the number of new cases in that LTLA-week, taken directly from the UK linelist, and only those that there were pillar 2 and PCR-confirmed. This is all 0s in the file I’ve given you because I’m not 100% sure you’d be allowed access. Probably no problem, but let’s use dashboard cases for now and we can swap to this nearer the end if we want to.

**prop\_travelling\_to\_work** – The proportion of people in this LTLA that reported themselves as travelling to work pre-pandemic (2011), as taken from nomis data (nomis\_2020\_07\_03\_095643.xlsx). Note that a couple of LTLAs didn’t have these data available so some suitable surrogates were used, we can talk more about that if it becomes necessary, mainly taking average of neighbouring areas.

**prop\_not\_in\_work** – The proportion of people (aged 16-74) in this LTLA that reported themselves as not in work pre-pandemic (2011) as taken from nomis data (nomis\_2020\_07\_03\_095643.xlsx).

**prop\_white\_british** – The proportion of the LTLA population that reports themselves as “White British” according to 2016 census data (EthnicitybyArea2016.xlsx). Note that this data is rounded to the nearest thousand.

**IMD\_Average\_score** – The average “Index of Multiple Deprivation” score according to the 2019 report (File\_10\_-\_IoD2019\_Local\_Authority\_District\_Summaries\_\_lower-tier\_\_.xlsx). *The higher the score, the more deprived the area.* This averages together measures for education, crime, health, barriers, living conditions, and others, you can dive in to the file to get the details.

**IMD\_Rank\_of\_average\_score** – Each LTLA is ranked, with 1 being the most deprived, and 317 the least deprived (devolved nations don’t have this info). This is also taken directly from File\_10\_-\_IoD2019\_Local\_Authority\_District\_Summaries\_\_lower-tier\_\_.xlsx

**resident\_earnings** - median weekly earnings (in £) of full-time workers according to lmregtabli01april2020.xlsx

**mean\_age** – This has been manually calculated from the data in MYE2 of “ukpopestimatesmid2020on2021geography.xlsx”, by summing all those from age 0 to 89, and dividing by the total population. Note that we do technically therefore ignore the small 90+ category in this.

**prop\_o65** – The proportion of this LTLA’s population that is over (*and including*) the age of 65, as calculated from the data in MYE2 of “ukpopestimatesmid2020on2021geography.xlsx”, by summing all those over and including the age of 65 and dividing by the total population.

**mean\_popden** – Similar to, but different from Pop\_per\_km2 above. This measure is also in average population per km^2, but is instead taken from the source “Copy of SAPE22DT11-mid-2019-lsoa-population-density.xlsx” This source has some max,min,standard deviation, so could be returned to if you wanted to capture some of the heterogeneity in pop\_den of the LTLA.

**Median\_annual\_income** - Data on the median annual income for each LTLA is taken from the 2021 and 2022 editions of the ONS Annual Survey of Hours and Earnings (ASHE) (<https://www.ons.gov.uk/employmentandlabourmarket/peopleinwork/earningsandworkinghours/datasets/placeofresidencebylocalauthorityashetable8>) . Earnings data is split by financial year; all weeks before April 6th 2021 use the 2020 earnings data, and all weeks after April 6th 2021 use the 2021 earnings data. I manually extracted these from the link above.

**no\_jobs** – Taken from the same link as the previous variable, this is an estimate of the number of jobs (in thousands) based in this LTLA during this year.

**retail\_and\_recreation\_percent\_change\_from\_baseline** – The first of the Google mobility measures, taken from: <https://www.google.com/covid19/mobility/> . I have manually averaged them over the week, as information is available daily. “The data shows how visitors to (or time spent in) categorized places change compared to our baseline days. A baseline day represents a normal value for that day of the week. The baseline day is the median value from the 5‑week period Jan 3 – Feb 6, 2020.” You can read their full explanation and details here: <https://support.google.com/covid19-mobility/answer/9824897?hl=en&ref_topic=9822927&sjid=14507744449389682797-EU>

In short, it’s the percentage change from a pre-pandemic baseline to the number of visits to retail and recreation sites in this LTLA for this week.

“The Residential category shows a change in duration—the other categories measure a change in total visitors.”

**Alpha\_proportion** – The proportion of sequenced cases from this NHS region for this week that came back as being “Alpha variant”. Note that I don’t have the denominator of tests available with this. This data was taken from the UK dashboard – “Portal\_variant\_data.csv”. Note that this did not capture the Alpha period very well, so I manually supplemented this with data from VAM (Variant and mutation) sources.

A key thing to note, this was NOT available at LTLA level, only the 9 NHS regions, so each LTLA in an NHS region will have the same proportion info here.

**s\_Wild\_prop** – This is an alternative metric of proportion based on SGTF (S-gene target failure) data. This IS available at the LTLA level. Source data not provided, as again it’s roots are in the linelist.

The UKHSA line list records the result of spike gene target failure (SGTF) for PCR-confirmed cases. This simplified test relies on the fact that the Alpha (B.1.1.7) and Omicron (B.1.1.529, BA.1, BA.4 and BA.5) variants of concern (VOC) share specific mutations in their spike gene not observed for the Delta variant or original wild-type. McMillen et al. (2022) report that at the peak of their corresponding wave, the positive predictive value of the SGTF was 98% for Alpha and 100% for Omicron. As such, we classify each case from the linelist into variants via the following criteria:

From the start of the pandemic to March 27th 2021, s-gene positive cases are tagged as “wild-type", s-gene negative cases are tagged as “Alpha variant".

From March 28th 2021 to September 11th 2021, s-gene positive cases are tagged as “Delta variant", s-gene negative cases are tagged as “Alpha variant".

From September 12th 2021 to January 1st 2022, s-gene positive cases are tagged as “Delta variant", s-gene negative cases are tagged as “Omicron variant".

From January 2nd 2022 onwards, we assume all cases are of the Omicron variant, as the BA.2 subtype is s-gene positive, and cases of Delta are seen to be negligible from this point on nationally.

From this classification, we record for each LTLA, weekly, what proportion of cases are made up of each variant.

**CCG\_2019\_Code** – Another way of dividing up geographies is by NHS “clinical care groups” (CCGs). These boundaries are how NHS funding are distributed. Note that these boundaries actually cut through some LTLA boundaries, so this code (and name) details the CCG that the majority of this LTLA belongs to.

**NHS\_registered\_population** – The number of registered NHS users in this CCG.

**Core\_services\_funding –** NHS funding given each year to each CCG is split across “core services”, “primary care”, and “specialised services”. This is the amount of “core services” funding (in £) for this financial year, to this CCG. Data taken from: <https://www.england.nhs.uk/allocations/previous/allocations-for-2019-20-to-2023-24/>, which I manually extracted in NHS\_funding\_allocations.csv.

**Core\_services\_funding\_by\_weighted** – This funding is allocated based on some maths that they do, they estimated a “weighted population” where they factor in the age, deprivation, and care needs of a population. This weighted population is in NHS\_funding\_allocations.csv. This variable is the Core Services funding amount / weighted population. This number is therefore a better indicator of “imbalances” in funding provision.

**Unringfenced** – These next 22 variables (unringfenced to tax\_income\_guarantee\_business\_rates) details all the COVID-specific “special” emergency funding that was given out to LTLAs, as taken from here: https://www.gov.uk/government/publications/covid-19-emergency-funding-for-local-government

Outbreak management details here: <https://www.gov.uk/government/publications/contain-outbreak-management-fund-2021-to-2022/contain-outbreak-management-fund-guidance-financial-year-2021-to-2022>

Specific details of restrictions on how the funds could be used can be found in the above links.

All amounts are in “£ millions”.

Unringfenced was free to be used as LTLAs saw fit. The main three I would focus on are “unringfenced”, “COMF” (Contain Outbreak Management fund), and “ASC” (Adult social care), although maybe there’s interesting things hidden in the rest!

**Centroid\_x** – You can ignore these. They are scaled x/y coordinates I generated to do some spatial maths with.