Initial Data Collection:

After settling on the ONS and DEFRA data, we downloaded these into a local file which we then performed the cleaning of the data on. The happiness data was all located in a singular sheet within an excel file which we downloaded in a csv format. The Pollution data had individual csv files for each year.

Our final goal was to have three data sets with comparable data for PM10, PM2.5 and Happiness means.

We used [numpy](https://numpy.org/) and [pandas](https://pandas.pydata.org/) for this.

**Happiness**

The happiness data was categorised by local authorities and counties within each region of the country.

We uploaded the Happiness means csv files, having different dataframes for each region. This would allow us to split the data between North and South as we had the indicative regions for each (NORTH AND SOUTH DIVIDE LINE REFERENCE). The column names needed to be consistent throughout all dataframes to make the data more easily readable and so we could merge and concactenate data by column names.

The happiness was then concactenated into north and south. All missing data was replaced with the integer, “0”. A key difference in the data cleaning here, compared to the workshop was that regex was False as we did not want the missing values to be considered substrings and replace Local Authority Names with this value.

We used listwise deletion (link to methodology?). Therefore, if there was a missing value in any column within a given row, the whole row was deleted.

The clean datasets were redownloaded as csv files to be uploaded to a shared group folder.

**PM10 & PM2.5**

The pollution data had many more data points than the Happiness data and multiple csv files which had to be combined into a singular dataframe. It was also categorised by x and y coordinates within the British National Grid system and a singular identifiable UK Grid Code.

We saved all the files for PM10 and PM2.5 into folders with their names respectively which we used as a global file path to automate the reading of each file within pandas. These were all concactenated into a dataframe, providing two dataframes- one for PM10 and one for PM2.5.

Once concatenated, the data was grouped by UK Grid Code. These were cleaned and any blank cells or missing datum cells were given the value of 0. Listwise deletion was completed for these and clean datasets for each variable were saved as csv files to make accessible to the whole group.

**WHAT NOW**

We were at a point were we had three datasets with no missing data. However, the pollution data still included data for Wales, Scotland and Northern Ireland and was not categorised by Local Authority (like the happiness data). At first we considered using external libraries such as [pyBNG](https://pypi.org/project/PyBNG/) or [pyProj](https://pyproj4.github.io/pyproj/stable/) to translate the coordinates into the WGS84 system which [openstreetmap](https://www.openstreetmap.org/#map=4/-15.13/-53.19) uses. However, pyBNG kept giving errors and the pyPROJ projections were not very accurate. When searching for software that could use the British National Grid (BNG) coordinate reference system (CRS), we came across [QGIS](https://qgis.org/en/site/), a free and open source software which is widely used and has a lot of documentation to support its usage. This meant our data could be uploaded using directly the x and y coordinates provided.

The clean datasets for pm10 and pm2.5 were uploaded to the software. The coordinates for each data point is the centroid of a 1km by 1km quadrant ([a](https://uk-air.defra.gov.uk/data/pcm-data)). For each coordinate point we used a 4 segment buffer with a square end cap each at a 500 metre distance from the coordinate.

We used the ONS shapefile ([b](https://geoportal.statistics.gov.uk/datasets/ons::lad-dec-2021-gb-bfc/explore?location=55.217379%2C-3.265847%2C6.87)) to have boundaries for each local authority.The data had a column with the “Area Code”. Upon [researching](https://geoportal.statistics.gov.uk/documents/coding-and-naming-policy-for-uk-statistical-geographies-1/explore) what these meant, we found that all codes starting with the letter “E” denominated England data.

We used the “join by location (summary)” function which used the ONS’ Local Authority bounds and found the mean of the data for each coordinate within that bound. This meant we had means for each local authority. Any features within this map with an area code that did not start with “E” was deleted.

We cross-referenced by merging layers and finding if there were any areas for which there were no matching data. This came up with 9 locations: PLS INCLUDE THE NAMES IN THE PIC I SENT IN GC. We now had a total of 300 locations for which we could compare data.

Now that there was only data for England, we were able to produce visualizations by using a graduated marking system for which we used automatic classification of the data by the QGIS software. This was repeated for each year, producing a visually comparable map with happiness and pollution data from 2011 to 2021.

**Why Listwise**

One of the biggest challenges in cleaning the data was deciding how we would approach the missing data. Would we attempt to use imputation?

Imputation can be done in many ways with some simple imputation techniques including linear regression.

Data for which all years have missing values (aka a row full of 0s) has an equal probability for any given year to have missing data (<https://stefvanbuuren.name/fimd/sec-simplesolutions.html> - go to MCAR (pls don’t use as a reference but just as a place to understand stuff from-- the sagepub thing has this as well).

Therefore, “Specifically, if the data are MCAR, then the reduced sample will be a random subsample of the original sample”- **sagepub**.

So listwise for that works bc all values are missing.

For singular missing values, there is enough complementary data that all local authorities for pollution still had means found for them but with different unique values to each (but all still had more than y (i’ll find this value asap) unique data points.

The best way to deal w pollution missing data as found by this study was to use MICE(multiple imputation chained equations) because they account for the uncertainty within each imputation however we would need much more data to have a trained dataset which could then give accurate estimations for missing values. This was outside of the scope of our study.

The next best way would be to run a KNN. However, listwise was satisfiable for pollution and using KNN for the happiness data set which is already in means and covers larger areas as they are categorised by local authority woul yield unreliable results and this was also outside of scope of our study. Therefore, we went for the slightly biased but still satisfiable solution. ([ref](https://ueaeprints.uea.ac.uk/id/eprint/77669/1/Accepted_Manuscript.pdf))

REFERENCES FOR DELETION:

LITERALLY the holy grail: has info on listwise, different imputation techniques etc. ---> amazing

<https://methods-sagepub-com.libproxy.ucl.ac.uk/book/missing-data/n3.xml>

Readings for this:

<https://onlinelibrary.wiley.com/doi/full/10.1002/0471264385.wei0204>

<https://www-sciencedirect-com.libproxy.ucl.ac.uk/science/article/pii/S1352231006006959>

<https://ueaeprints.uea.ac.uk/id/eprint/77669/1/Accepted_Manuscript.pdf>

Key features of choropleth maps:

* Blank areas are local authorities with missing data. These are: Boston, Oadby and Wigston, North Northampton, West Northampton, Gravesham, Adur, Isles of Scilly and Richmondshire.
* The boundaries for the pollution data have decreased. The highest PM10 value in 2011 was 25.3 whilst in 2021 this was 17.29. This decrease is even more clear in the PM2.5 where the maximum value was 17.64 in 2011 whilst in 2021 it was 9.49.
* For Happiness, the upper bound has been mostly constant with less than a 0.2 variation throughout the years.
* Greater London, South East England, and Yorkshire and The Humber have consistently shown higher levels of pollution compared to other England regions. Areas with most pollution tend to be in metropolitan areas.
* An outlier to this trend is in the Yorkshire and The Humber region directly below the Humber Estuary where the port of Grimsby & Immingham is located. In terms of cargo volume, this was the largest port in the UK between 2011-2019 and has been the second largest since 2020. This may account for the higher emissions shown in a less densely populated area. ([a](https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/9258/port-freight-statistics-full-summary-2011.pdf),[b](https://www.gov.uk/government/statistics/port-freight-statistics-2012-final-figures),[c](https://www.gov.uk/government/statistics/port-freight-statistics-2013-final-figures),[d](https://www.gov.uk/government/statistics/port-freight-statistics-2014-final-figures),[e](https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/555338/port-freight-statistics-2015.pdf),[f](https://www.gov.uk/government/statistics/port-freight-statistics-2016-final-figures#:~:text=unitised%20traffic%20rose%20by%202,1%25%20to%204.5%20million%20units),[g](https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/762200/port-freight-statistics-2017.pdf),[h](https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/826446/port-freight-statistics-2018.pdf),[i](https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/908558/port-freight-statistics-2019.pdf), [j](https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1014546/port-freight-annual-statistics-2020.pdf),[k](https://www.gov.uk/government/statistics/port-freight-annual-statistics-2021/port-freight-annual-statistics-2021-overview-of-port-freight-statistics-and-useful-information))