Task 6.1: Sourcing open Data

Data source:

Summary:
Data on greenhouse gas emissions and removals, sent by countries to UNFCCC and the EU Greenhouse Gas Monitoring Mechanism (EU Member States) from the European Environment Agency. Homepage: https://www.eea.europa.eu

-This is an external data source from the European Environment Agency and is likely to be trustworthy.

-Usage data which has been tracked automatically by computer automated systems on a yearly basis. The time lag is unknown, but the time lag between data being recorded and becoming available for analysis is 2 years according to the information on the database.

European Environment Agency, '[DEPRECATED] Trends in emissions of greenhouse gases (IPCC sector classification)', accessed 2023-05-23, http://data.europa.eu/88u/dataset/data_trends-in-emissions-of-greenhouse-gases-ipcc-sector-classification-5

Data contents: The data contains monthly GHG emissions counts in European countries from X and Y. The counts are broken into X categories. What variables are included?

Data relevance: Historical trends often mirror upcoming trends. For this reason, this data can be used to predict future GHG emissions levels for planning, mitigation and adaptation purposes. The included geographical location data can help to illuminate vulnerable populations or areas that require additional mitigation or adaptation measures.

Limitations: The data is deprecated, meaning its failings are recognised. (expand) Potential biases in the data are avoided because they have been collected and distributed by the European Environment Agency, rather than countries themselves, which may have an incentive to alter or omit data in some way. Is it collected infrequently? Could it contain manual errors?

Explanation: This data set has been chosen due to my interest in Climate-related issues and sustainability. The source is collected on a geographical scale across Europe, which is where I live, and is directly relevant to me. As a trainer and educator in the Climate Crisis, I am interested in seeing what this data looks like in its raw form, and being able to carry out some predictive analysis of my own, which will deepen my understanding of the methodology used behind the collection of the data which I cite in my training. For my portfolio this is useful, because it fits in with the focus of my future employment role.

Data Profile

Info on the raw data: Shape: (298936,12)

Data Types:

```
0   Country_code       298936 non-null  object

 1   Country           298936 non-null  object
 2   sort                298936 non-null  int64

 3   Pollutant_name      298936 non-null  object

 4   Format_name         298936 non-null  object

 5   Sector_code         298936 non-null  object

 6   Parent_sector_code  295002 non-null  object

 7   Sector_name         298936 non-null  object

 8   Year                298936 non-null  int64

 9   emissions           298936 non-null  float64

 10  Unit                298936 non-null  object

 11  Notation            141893 non-null  object
```

Data Statistics:

```
Sort emissions

count298936.0000002.989360e+05

mean20.5622272.689009e+03

std18.4718476.902170e+04

min1.000000-4.475200e+05

25%3.0000000.000000e+00

50%4.0000000.000000e+00

75%39.0000000.000000e+00

max42.0000004.128192e+06
```

Action taken:

#dropping unnecessary columns in an adjusted database, keeping the original: 1. Format_name (same entry for all rows, adds nothing to the data), 2. Notation (not necessary for analysis)

#changing data type 'emissions' from 'float64' to 'int64'

#replacing missing values: All sector values of '0' have a corresponding value of 'NaN' in the parental sector. This will be replaced with '0'

#filling missing values with '0' to match info in the 'sector code' column (i.e all parent sector code

Preliminary findings:

```
Highest sectors are: Other                           40312

Total National Emissions and Removals     3934

CO2 Emissions and Removals from Soil      3934

Rice Cultivation                          3934
```

Agricultural Soils	3934

Limitations and ethics:

The major limitation of this database is the age. More recent contributions are not currently accessible. Other issues include the fact that this data is presented by each country for assessment rather than collected independently, so collection methodology may differ, or collection may be incomplete/skewed as it is a collection of data from many different sources.

Questions to explore.

-Which sectors have the highest emissions rates?

-How have emissions rates in each area changed over time, and what might predictions for the future look like?

-In which sectors/countries has progress been made/not been made?