The Stack Overflow developer survey for 2021 is a collection of responses from users of the Stack Overflow website. It contains different kinds of fields about the users and their occupations. The data will require some cleaning as not all fields are filled for every respondent. This dataset does not contain much numeric data and will require a large amount of cleaning and processing which makes visualisation difficult.

The Breast Cancer database contains data about women suffering from breast cancer diagnosed between 2006 and 2010 from the National Cancer Institute in the US. The data contains fields about the patients like age, marital status, whether they were alive or not at the time of data collection (2017). It also contains medical information about the patients. This data set does not seem like it will require much cleaning but contains good numerical data for visualisation. It also contains medical information that will provide an interesting challenge to learn about and understand. For these reasons we picked this dataset to work with.

The CO2 emissions data contains data about the CO2 emissions of countries between 1750 and 2022. It contains many 0's from countries in years where data has not yet begun to be collected. It does not contain as many fields as the other datasets.