#DataMadness Project Proposal Form

i-number	Student name
6256590	Aurélien Bertrand
6260490	Spencer Olson
6260219	Oscar van Bommel Barcia

(Provisional) Title:
Can change during the project, but make sure that you pick a title that attracts the audience

The secrets behind European air quality data.

Questions and Objectives:
List at least 3 questions (high-level) of what you want to address in your analysis. More are welcome, but think of the workload

1. Do bigger countries have worse air quality?

2. Is there a correlation between the altitude and the type of pollutant?

3. Do the regulations introduced by the European Commission change anything compared to other countries?

The Dataset
Attach the dataset or point to a web link that is available (if very big or already existing). Highlight why it can answer your questions. If the dataset does not come with datasheets, you are going to create them to best of your knowledge (see extra form)

Link to the dataset: https://discomap.eea.europa.eu/App/AirQualityStatistics/index.html

This dataset provides information about the air quality in European countries. Records from 1969 up to 2023 is provided. Which pollutant is present in which quantity is provided for each year and location.

Rough Outline of the Analysis
This is a rough sketch of the analysis you want to do on the dataset and which techniques you should use to answer your questions. Not detailed at the proposal, but should act as a guide as you work on the project.

First, a general analysis of the dataset will be performed. This includes cleaning and potential translations to apply to the data. Next, relevant information will be kept, and the rest discarded.

Next, an ethical analysis will be performed, defining any potential privacy concerns arising with the collection of the data.

In order to answer the first question, we will check for potential correlation between country size and air quality, looking at which pollutants are present in each country and in which quantity. To do so, we will statistically test for any correlation and use visualization.

To answer the second question, we will check for potential correlation between the altitude and the kind of pollutant observed. Again, statistical tests will be used together with a visualization.

Finally, we will analyze trend over time and check what years where the best in terms of air quality for each country. We will also check if the pollution gets flattened over time, since the regulations were introduced. This will be used to answer the last question. Together with this, we will use regression to try to forecast the air quality in future years for each European country. This could give information whether further regulations should be added in Europe.