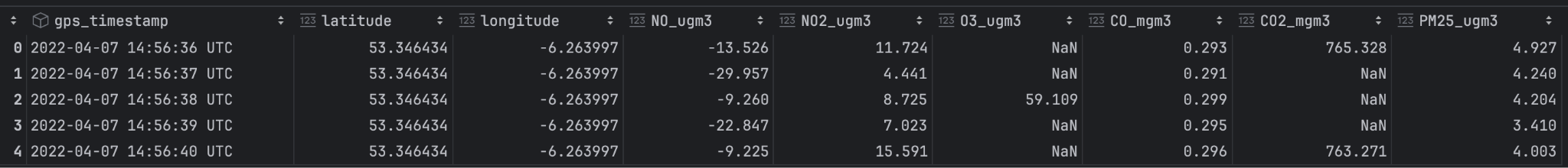
ML Project Memo

**Projectname/ title:** Hyperlocal Air Quality Prediction in Dublin

**Group Members:** Elena Dreyer, Giulia Petrilli, Killian Conyngham

### Dataset Description

* The data originates from a collaboration between Google and Dublin City Council
* Measurements were taken by Google's electric Street View car, which was equipped with Aclima’s mobile air sensing platform
* Data collection occurred mainly on weekdays (Monday–Friday) from 9:00 am to 5:00 pm, representing typical daytime air quality between **May 2021 and August 2022**
* The Street View car traveled at normal traffic speeds and followed traffic flow, recording data street by street and measured air quality at **one-second intervals**
* Pollutants measured include Carbon Monoxide (CO), Carbon Dioxide (CO2), Nitrogen Dioxide (NO2), Nitric Oxide (NO), Ozone (O3), and Particulate Matter PM2.5
* Source: [Aclima & Google 2022](https://data.smartdublin.ie/dataset/google-airview-data-dublin-city)



### What are your units of analysis?

Each row in the dataset represents an individual measurement of air quality indicators (such as concentrations of NO, NO2, O3, CO, CO2, and PM2.5) associated with a specific GPS location (latitude and longitude). In our dataset the measurements are time specific (in seconds), but we plan to average over a larger time period for each location (daily). Therefore, our unit of analysis would be the air quality measurements at a specific location.

### 

### What are you trying to predict

Our label will be the **street level pollution level in Dublin**. Based on the dataset, we will focus on one pollutant (e.g. NO2, PM2.5) and set it as our label. Our project will take the averaged pollution data, and for new locations, predict the pollution level. The aim is for these predictions of untracked roads to be better than the simple average between the pollution level of the two nearest roads. Our project applies different models to the data, and we envision for the first one to be a regression.

### What will you use to make that prediction?

We use **past pollutant measurements** (PM2.5, NO2, CO2), depending on how complete the data for each pollutant is. PM2.5 is a widely recognized measure, choosing it would probably be of most relevance. We plan on using **weather** and **traffic** data too.

We are planning on merging different sources of data, on weather, on traffic, and on air pollution, and which share the location coordinates, hence all the inputs and output variables are specific to location.

### How would this prediction be used in a decision-making context?

To get more location specific information regarding air pollution can be very important for governments in considering the health outcomes of their populations. As exposure to pollutants takes place on the individual level, even cities with air quality measures deemed to be safe on average may still contain neighbourhoods or streets where individuals are consistently exposed to unsafe levels of pollutants. This risk is especially pronounced for individuals living near busy roads or large point sources of pollution such as industry. Quality predictions of where air quality deviates from safe levels can help inform decisions on targetted mitigation measures, industrial zoning and planning permission, as well as transport infrastructure planning and future health expenditure forecasting. Individuals in high risk groups for adverse effects of pollutants may also find this information useful as a factor when considering where to work and live.

### Miscellaneous

In our limitation section we will acknowledge that the data was collected during Covid, and that lockdown might have influenced the pollution levels and the traffic. Ideally, our analysis would be the base for a website that outputs pollution estimates as a function of time and space. However, we will first focus on the machine learning aspect of the project and later ideally expand on software development.