1. Find dataset(pollution, public health)
2. Purpose (find the relationship between pollution and public health)

* Find which chemical affects which diseases (+how much?)

1. Wrangling & Analysis

* **#Pollution(Air) (1979-2019):**
* Choose columns: sp\_name, sample\_datatime, param\_id, PV
* Remove data Non-related to pollution ex) sws, vwd, vws, dbt
* Data we using: API, BPM2.5, CO, NO2, O3, PM10, SO2
* Group by Region, Month
* Make average every 4years because of death dataset every 4 years
* **#avoidable death (2008-2016 by 4 years):**
* Data we using:IIschaemic heart disease, Colorectal cancer, Suicide and self-inflicted injuries, COPD, Cerebrovascular diseases, Breast cancer, Transport accidents, Accidental poisoning by and exposure to noxious substances, Diabetes, Skin cancer Prostate cancer, Falls, Kidney cancer, Selected invasive infections,  Heart failure, Rheumatic and other valvular heart disease

Notes:

1. We can investigate the trend of air pollutants over the years
2. Do we need to normalize our data?