**Maria Battle, Lab Book Notes**

Dataset OpenAQ\_07262018

The OpenAQ\_07262018 dataset contains air quality metrics collected on sensors July 26, 2018. The dataset was downloaded on July 27, 2018 from the OpenAQ website using “Download Yesterday’s Data”.[[1]](#footnote-0)

**Lab Book Index**

1. Data Element Dictionary (DED).......................p.2
2. Data Quality……………………………………...p.3
3. Data Cleaning……………………………………p.4
4. Data Transformation…………………………….p.5
5. Exploratory Data Analysis Ideas……………….p.6

**Open Source, Unofficial Data Element Dictionary[[2]](#footnote-1)**

* **PM** - particulate matter, a term for a mixture of solid particles and liquid droplets found in the air. Some particles, such as dust, dirt, soot, or smoke, are large or dark enough to be seen with the naked eye. Others are so small they can only be detected using an electron microscope.
  + PM10 : inhalable particles, with diameters that are generally 10 micrometers and smaller; and
  + PM2.5 : fine inhalable particles, with diameters that are generally 2.5 micrometers and smaller.
  + How small is 2.5 micrometers? Think about a single hair from your head. The average human hair is about 70 micrometers in diameter – making it 30 times larger than the largest fine particle.
* **SO2** -sulfur dioxide
* **NO2** - nitrogen dioxide
* **CO** - carbon monoxide
* **BC** - black carbon
* **O3** - ozone

Unit Definitions[[3]](#footnote-2)

* **ppm** - parts per million
* **µg/m3** - The concentration of an air pollutant (eg. ozone) is given in micrograms (one-millionth of a gram) per cubic meter of air or µg/m3.

**Data Quality Notes**

* Values in data sample appear to not always match the website map.
* Parameters appear inconsistent. Recommend that units of measure should be separately analyzed and visualized.
  + bc, co, no2, o3, so2, o3 all have a µg/m3 and ppm units of measure.
  + pm10 and pm25 only have µg/m3 and no a ppm unit or measure.
* The OpenAQ website visualization appear to only use UTC time.
* In the data sample, local date and time appear to have two different value formats. Example: 14:00:00-10:00, 00:00:00+03:00.
* Data sample has negative values which need interpretation.
* Data sample has a possible invalid value of “9999999” (isolated to TR for the July 26 set).
* A Data Element Dictionary (DED) or overview guide has not been provided for fields and values.
* The OpenAQ website provides a heat-map type cluster analysis of similar parameter values based on the last sensor reading (UTC) in that area. Reasoning for thresholds is unknown.
* Multiple parameters are collected during one 15 - 60 minute sensor time cycle, for a daily duration of anywhere between 00:00 to 23:15 UTC on July 26, 2018. Parameter units (ppm or µg/m3) vary by region. The reason for this is unknown, and the researcher should investigate if this indicates that data is collected or measured differently across regions.
* We don’t have information about how the sensors work.

**Data Cleaning Notes for This Dataset** - all deleted data has been saved in separate files.

* Attribution column was deleted. E.g. values for row 1: [{"name":"SINCA","url":"http://sinca.mma.gob.cl/"},{"name":"Electroandina S.A."}]
* 148,785 rows removed for IL (Israel) - character translation unreadable in R and Excel
* 50 rows removed for MN (Mongolia) - character translation unreadable in R and Excel
* 84 rows removed for TR (Turkey) - character translation unreadable in R and Excel
* Some negative values and “9-type” values were kept in the dataset experimentally. Meaning was requested but not been provided. They seem to be a misreported value and skew the data. They may be isolated to a few countries.
* 14 rows deleted for Ankara and Trabzon Turkey (TR) with 9999999 values
* 328 rows for AU, AE, BH, ET, NL, UG, IN, LK, BH, UG deleted with -999 or -9999 values.

**Data Transformation Notes for This Dataset**

Parameters bc, co, no2, o3, so2, o3, pm10 and pm25 we parsed into separate columns.

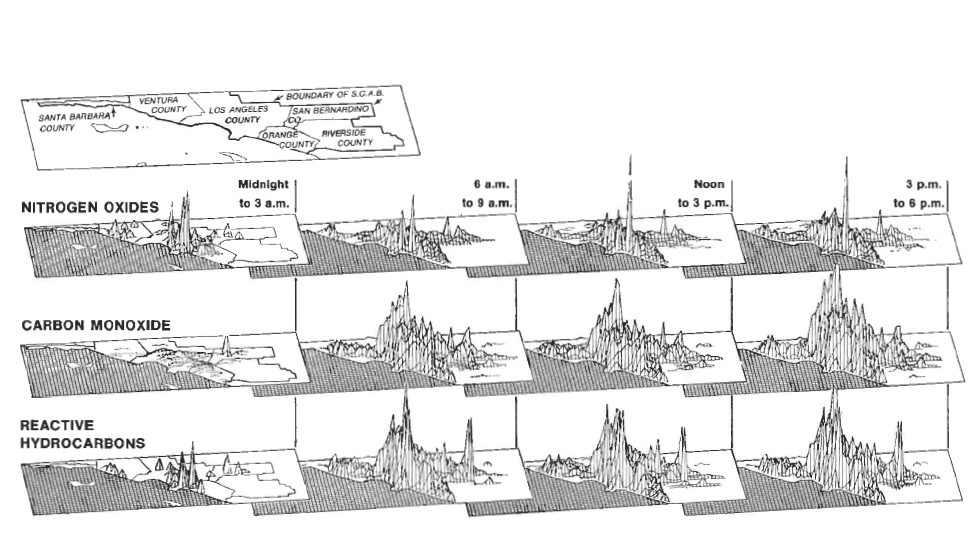
* 1=has a value for this parameter
* 0=does not have a value for this parameter

All 49 countries in this dataset were renamed as numerical values to make it more flexible. Country Legend:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| AD=1 | CO=11 | HU=21 | MN=31 | SI=41 |
| AE=2 | CZ=12 | ID=22 | MT=32 | SK=42 |
| AU=3 | DK=13 | IE=23 | MX=33 | TH=43 |
| BA=4 | ES=14 | IN=24 | NL=34 | TR=44 |
| BD=5 | ET=15 | KW=25 | NO=35 | UG=45 |
| BH=6 | FI=16 | LK=26 | NP=36 | US=46 |
| CA=7 | FR=17 | LT=27 | PE=37 | VN=47 |
| CH=8 | GB=18 | LU=28 | PT=38 | XK=48 |
| CL=9 | HK=19 | LV=29 | RS=39 | ZA=49 |
| CN=10 | HR=20 | MK=30 | SE=40 |  |

**Exploratory Data Analysis Ideas** - none of these have been requested by the customer. They are just a personal brainstorm.

* Correlation of weather patterns
* Correlation of the history of coroner reports for cause of death
* Correlation of inland or coastal location
* Correlation of number of sensors in a related region
* Correlation of sensor proximity to roads, factories or other commercial industry buildings (e.g. airports and times of heaviest fight activity), and whether airflow is a sensor factor. Add the street location of the sensor based on coordinate search in Google Maps. This may be scoped to areas that are >85.6 µg/m3 and correlated with the distance between the sensor and other industrial buildings. Google Maps will allow you to measure if you right click on the map and select “measure distance”.
* Summarize parameter values variance over an approximate 24 hour period for a city, state or country. Example by Tufte in his book, The Visual Display of Quantitative Information.



* Develop a dB of aggregate samples based on certain parameters.regions <https://openaq-data.s3.amazonaws.com/index.html>

1. https://openaq.org/#/locations?\_k=6v3rq9 [↑](#footnote-ref-0)
2. Sources:

   <https://www.epa.gov/pm-pollution/particulate-matter-pm-basics>

   <https://www.eea.europa.eu/themes/air/air-quality/resources/glossary/g-m3> [↑](#footnote-ref-1)
3. https://www.epa.gov/ [↑](#footnote-ref-2)