

Project Proposal: Sofia Air Quality Analysis

Group Members

Yukun Fang (yfang67), Mengkun Chen (mchen373), Mengqi Li (mli653), Runze You (ryou3)

Data Description

URL: <https://www.kaggle.com/hmavrodiev/sofia-air-quality-dataset>

Soifa air quality dataset describes daily five air-quality measurements-PM2.5, PM10, pressure, tempreture and humidity- from July 2017 to July 2019.

```
bme1707<-read.csv("data/2017-07_bme280sof.csv", header = TRUE)
bme1707<-bme1707[, -1]
bme1707$timestamp<-sub("T.*$", "", bme1707$timestamp)
head(bme1707)
```

##	sensor_id	location	lat	lon	timestamp	pressure	temperature	humidity
## 1	2266	1140	42.738	23.272	2017-07-01	95270.27	23.46	62.48
## 2	2292	1154	42.663	23.273	2017-07-01	94355.83	23.06	59.46
## 3	3096	1558	42.700	23.360	2017-07-01	95155.81	26.53	44.38
## 4	3428	1727	42.624	23.406	2017-07-01	94679.57	28.34	38.28
## 5	3472	1750	42.669	23.318	2017-07-01	94327.88	26.31	46.37
## 6	1952	976	42.709	23.398	2017-07-01	95314.52	22.66	56.55

```
sds1707<-read.csv("data/2017-07_sds011sof.csv", header = TRUE)
sds1707<-sds1707[, -1]
sds1707$timestamp<-sub("T.*$", "", sds1707$timestamp)
head(sds1707)
```

##	sensor_id	location	lat	lon	timestamp	P1	P2
## 1	753	361	42.626	23.378	2017-07-01	13.77	6.80
## 2	1022	500	42.637	23.332	2017-07-01	13.33	7.73
## 3	2265	1140	42.738	23.272	2017-07-01	25.33	6.57
## 4	2291	1154	42.663	23.273	2017-07-01	15.07	9.67
## 5	3095	1558	42.700	23.360	2017-07-01	15.60	6.43
## 6	3427	1727	42.624	23.406	2017-07-01	13.73	6.43

Variables:

sensor_id: meteorological sensors reporting air-quality measurements

location, lat, lon: geographic information of sensors around Sofia

timestamp: observation dates (originally specified to seconds)

P1, P2: coarse particles measurements, representing PM2.5 and PM10 respectively

pressure, temperature, humidity: meteorological measurements

Statistical Problems

For this dataset, we try to solve the problems as follows:

1. Are there any trends in pressure, temperature, humidity, PM2.5 and PM10?
2. Is there any relationship between pressure, temperature, humidity and PM2.5/PM10?
3. How do pressure, temperature and humidity influence PM2.5 and PM10 respectively?

Statistical Methods

We intend to use time series plot to figure out the trend for these variables. Plus, we plan to use linear regression to interpret the relationship between pressure, temperature, humidity and PM2.5/PM10.

Computation Plan

We plan to use CHTC to do parallel computation for the .csv files of each month. Then we merge the small datasets into one large dataset and do statistical analysis.