**데이터셋을 확률/통계와 함께 일반화하기**

공장 대기모니터링은 공장으로부터 발생하는 대기오염의 정도를 측정하는 기술이다. 공장 대기모니터링은 고정오염원 대기모니터링이라고 불리우기도 하며, 여기서 고정오염원은 제조공장, 플랜트, 화학공장 등을 의미한다. 여러분이 알고 있는 미세먼지 등을 측정하는 기술은 국가대기질 모니터링이라고 하며, 다음과 같이 고정오염원 대기 모니터링과 차이가 있다.

* 고정오염원 대기 모니터링은 상대적으로 실시간 모니터링을 요구한다. 그 이유는 만약 일정량 이상의 대기오염 물질이 배출되면 벌금이 부과되기 때문이다.
* 고정오염원 대기 모니터링은 일정한 패턴을 가질 확률이 높다. 그 이유는 공장의 가동 과 배출되는 오염물질의 양이 연관되어 있기 때문이다.

다음은 고정오염원 대기 모니터링 시스템에 의해 수집되는 측정 메시지의 스키마이다.

|  |
| --- |
| Message schema {  **datetime** timestamp; // ISO 8601 format in UTC  **float** FM; // flow metering (cc per minute: ccm)  **float** NOX; // nitrogen oxide (parts per million: ppm)  **float** O2; // oxygen (ppm)  **float** temp; // air temperature (F’)  } |

첨부된 CSV 파일은 위의 스키마를 따르는 1달동안(3월) 수집된 데이터셋이다. 학생들은 다음 요구사항을 수행해야 한다.

1. csv 데이터를 가져오기 및 데이터를 메모리로 불러오기.

- 데이터의 수는 1000개 이상이어야 한다

- descriptive statistics 및 histogram을 적용하여 데이터셋의 정보를 파악해야 한다.

- 값가 없는 경우(Nan, Null), 전후 값의 평균치를 넣는다.

2. Datatime의 period 기능을 사용하여 주어진 데이터를 7일단위로 그룹화한다.

3. 특정 컬럼(columns)에 대해 10개, 100개, 1000개 bins을 나누고 Probability Mass function(PMF)를 구하여라

- pd\_Series.value\_counts(normalize=True, bins=10)를 사용하면 구성할 수 있다.

4. 세 개의 임의 컬럼 x, y, z이 PMF px, py, pz로 변환되었다고 가정하다. 이때 결합분포(Joint distribution) pxyz를 구하여라.

- 중첩 for-loop를 사용하여 구하시오.

- chaintools나 itertools 라이브러리를 사용하여 pxyz를 구하시오.

5. px의 누적함수분포(CDF) cx를 구하시오

- itertools의 acculmulate 함수를 이용하여 cx를 계산하시오. (이때, pmf는 오름차 순으로 이미 정렬되어 있어야 한다

**Further discussion (추가점수: 문제당 1점)**

6. 현재 데이터셋의 특정 컬럼의 평균과 분산을 계산하고, 이로부터 Gaussian distribution(=Normal distribution)을 사용하여 데이터를 재생성해보시오

7. 재생성된 데이터와 현재 데이터셋을 시간에 따른 측정값의 그래프을 출력하였을 때, 두 그래프의 모양이 서로 비교하시오. 또한, 만약 서로 비슷한 그래프를 생성하기 위해, 어떠한 작업을 추가로 진행해야할 지에 대해 기술하시오