빅데이터 프로그래밍
#3차 과제 – 대기질과 환경성 질환 분석

School of Information Convergence
Prof. Dong-Hyuk Im

광운대학교
KwangWoon University

전국 대기질 지수_데이터

	A	B	C	D	E	F	G	H
1	지역	측정일시	SO2	CO	O3	NO2	PM10	PM25
2	강원 강릉시	2021-01-01	0.0031	0.3	0.029	0.0157	19	12
3	강원 강릉시	2021-01-02	0.0034	0.4	0.029	0.0163	30	12
4	강원 강릉시	2021-01-03	0.003	0.4	0.036	0.0088	27	12
5	강원 강릉시	2021-01-04	0.0039	0.4	0.032	0.0165	22	12
6	강원 강릉시	2021-01-05	0.0039	0.4	0.031	0.0168	24	12
7	강원 강릉시	2021-01-06	0.0031	0.3	0.034	0.0129	13	12
8	강원 강릉시	2021-01-07	0.0032	0.4	0.031	0.0101	25	12
9	강원 강릉시	2021-01-08	0.0035	0.4	0.033	0.0101	11	12
10	강원 강릉시	2021-01-09	0.0039	0.5	0.034	0.0102	14	12
11	강원 강릉시	2021-01-10	0.0054	0.5	0.03	0.0131	18	12
12	강원 강릉시	2021-01-11	0.0037	0.5	0.026	0.0169	18	12
13	강원 강릉시	2021-01-12	0.0042	0.5	0.028	0.017	31	12
14	강원 강릉시	2021-01-13	0.0034	0.4	0.036	0.0152	118	12
15	강원 강릉시	2021-01-14	0.0038	0.5	0.027	0.0233	55	12
16	강원 강릉시	2021-01-15	0.0042	0.4	0.031	0.0215	76	12
17	강원 강릉시	2021-01-16	0.0026	0.5	0.031	0.0148	81	12
18	강원 강릉시	2021-01-17	0.0026	0.4	0.036	0.0079	17	12
19	강원 강릉시	2021-01-18	0.0032	0.4	0.033	0.0121	31	12
20	강원 강릉시	2021-01-19	0.003	0.4	0.028	0.0179	26	12
21	강원 강릉시	2021-01-20	0.0046	0.5	0.03	0.0258	63	12
22	강원 강릉시	2021-01-21	0.0034	0.2	0.038	0.0174	22	12
23	강원 강릉시	2021-01-22	0.0031	0.3	0.029	0.0188	17	12
24	강원 강릉시	2021-01-23	0.0025	0.3	0.038	0.0093	8	12
25	강원 강릉시	2021-01-24	0.0025	0.3	0.033	0.0125	20	12

환경성 질환 의료이용정보_데이터

	A	B	C	D	E	F	G	H
1	시도명	시군구명	요양개시연	요양개시월	질환	성별	연령군	진료합계
2	서울특별시	종로구	2021	1	비염	남자	0-5	50
3	서울특별시	종로구	2021	1	비염	남자	06-11	56
4	서울특별시	종로구	2021	1	비염	남자	12-17	36
5	서울특별시	종로구	2021	1	비염	남자	18-44	270
6	서울특별시	종로구	2021	1	비염	남자	45-64	249
7	서울특별시	종로구	2021	1	비염	남자	≥65	188
8	서울특별시	종로구	2021	1	비염	여자	0-5	49
9	서울특별시	종로구	2021	1	비염	여자	06-11	43
10	서울특별시	종로구	2021	1	비염	여자	12-17	42
11	서울특별시	종로구	2021	1	비염	여자	18-44	344
12	서울특별시	종로구	2021	1	비염	여자	45-64	303
13	서울특별시	종로구	2021	1	비염	여자	≥65	225
14	서울특별시	중구	2021	1	비염	남자	0-5	42
15	서울특별시	중구	2021	1	비염	남자	06-11	53
16	서울특별시	중구	2021	1	비염	남자	12-17	19
17	서울특별시	중구	2021	1	비염	남자	18-44	225
18	서울특별시	중구	2021	1	비염	남자	45-64	185
19	서울특별시	중구	2021	1	비염	남자	≥65	149
20	서울특별시	중구	2021	1	비염	여자	0-5	47
21	서울특별시	중구	2021	1	비염	여자	06-11	38
22	서울특별시	중구	2021	1	비염	여자	12-17	23
23	서울특별시	중구	2021	1	비염	여자	18-44	283
24	서울특별시	중구	2021	1	비염	여자	45-64	211
25	서울특별시	중구	2021	1	비염	여자	≥65	162

3차 과제 공지

- 대기질과 환경성 질환 분석

1) 대기 오염 물질과 질환 종류로 샘플 데이터를 만들어서 진료 합계 예측
2) 각 성별 진료비율 히스토그램 그리기
3) 대기 오염 물질과 질환, 진료합계를 이용하여 미성년 환자의 성별 분석 모델 구축 및 평가

*미성년자: 18세 미만(해당 데이터 기준)
*두 데이터의 지역과 날짜를 통일하여 활용
*반드시 문제마다 두 데이터를 전처리하여 하나의 데이터프레임으로 결합 후 과제 해결
*1번 문제 OLS 회귀 이용
*2번 문제 각 지역과 달별 총 진료 수(남성 진료 수 + 여성 진료 수)로 나누어 비율 산정
*3번 문제 데이터셋 분리(훈련/테스트) 매개변수: test_size=0.3, random_state=0
*3번 문제 평가지표: 정확도, 정밀도, 재현율, F1 스코어, ROC 기반 AUC 스코어
*1번, 3번 문제 질환 데이터(범주형)를 수치형으로 변환하여 회귀 분석
*제공된 csv의 데이터 값을 임의로 변경 X(데이터 내 결측치 없음)

- 배점: 15점
- 제출기한 : 06/17, 월요일 (밤 12:00까지)
- KLAS에 '학번_이름.ipynb' 파일로 제출(3개의 프로그램을 모두 1개의 ipynb에 따로구현)

KWANGWOON UNIVERSITY

1번 모델 구축 예시

```
1 Rformula = '진료합계 ~ SO2 + CO + O3 + NO2 + PM10 + PM25 + 질환
2 regression_result = ols(Rformula, data=data).fit()
3 regression_result.summary()
```

전국 대기질 지수_데이터

지역	측정일시	SO2	CO	O3	NO2	PM10	PM25
강원 강릉시	2021-01-01	0.0031	0.3	0.029	0.0157	19	12
강원 강릉시	2021-01-02	0.0034	0.4	0.029	0.0163	30	12
강원 강릉시	2021-01-03	0.003	0.4	0.036	0.0088	27	12

환경성 질환 의료이용정보_데이터

시도명	시군구명	요양개시연	요양개시월	질환	성별	연령군	진료합계
서울특별시	종로구	2021	1	비염	남자	0-5	50
서울특별시	종로구	2021	1	비염	남자	06-11	56
서울특별시	종로구	2021	1	비염	남자	12-17	36

예시와 같이 두 데이터를 결합하여 회귀 모델 구축

KWANGWOON UNIVERSITY

2번 데이터 예시

	시도명	시군구명	측정일시	성별	진료비율
0	강원도	강릉시	2021-01	남자	0.465625
1	강원도	강릉시	2021-01	여자	0.534375
2	[illegible]				
3	[illegible]				
4	[illegible]				
...	...	...	...	...	...
5491	[illegible]				
5492	[illegible]				
5493	[illegible]				
5494	충청북도	충주시	2021-12	남자	0.460115
5495	충청북도	충주시	2021-12	여자	0.539885

- 지역과 측정일이 같은 두 데이터의 진료 비율 합은 1
- 데이터 내 모든 지역과 시간을 전처리 후 히스토그램 시각화

KWANGWOON UNIVERSITY

3번 모델 구축 및 평가 예시

1. 데이터 내에서 x와 y 지정

```
X = data[['SO2','CO','O3','NO2','PM10','PM25','질환','진료합계']]
y = data['성별']
```

2. 실제 값과 예측 값을 이용하여 평가

```
accuracy = accuracy_score(y_test, y_pred)
precision = precision_score(y_test, y_pred)
recall = recall_score(y_test, y_pred)
f1 = f1_score(y_test, y_pred)
roc_auc = roc_auc_score(y_test, y_pred)

print('정확도: {0:.3f}, 정밀도: {1:.3f}, 재현율: {2:.3f}, F1: {3:.3f}'.format(accuracy,precision,recall,f1))
print('ROC_AUC: {0:.3f}'.format(roc_auc))
```

```
정확도: [redacted], 정밀도: [redacted], 재현율: [redacted], F1: [redacted]
ROC_AUC: [redacted]
```

KWANGWOON UNIVERSITY