수원대학교 데이터과학 경진대회

-국민건강영양조사 데이터분석-

주최: 수원대학교 데이터과학부

후원: 수원대학교 데이터연구소, DS&ML 센터



목차

- ▶ 경진대회 개요
- > 경진대회 데이터 소개
- ▶ 경진대회 데이터 생성
- > 경진대회 데이터 설명
- ▶ 경진대회 세부 규칙

▶ 경진대회 개요

경진대회 목표



►국민건강영양조사 데이터를 분석하고 결과를 제출함

경진대회 참가 방법

- ▶ 팀 구성
 - ▶ 수원대 재학생에 대해 4명 이하로 팀 구성
- ▶ 참가신청(단톡방 참가)
 - ▶ https://open.kakao.com/o/g8gesS0e (참여코드:data)
 - ▶ 단톡방 참가로 참가 신청을 대신함
- ▶ 국민건강영양조사 데이터셋 생성, 분석 및 결과 제출
 - ▶ 국민건강영양조사 사이트에서 데이터를 다운로드 받고, 데이터 프로세싱 코드를 실행하여 데이터셋을 생성함(자세한 데이터셋 생성 방법은 다음 슬라이드 참고).
 - ▶ 생성한 데이터셋을 분석하고, 결과를 이메일로 제출함

▶ 경진대회 데이터 소개 (국민건강영양조사 데이터)

국민건강영양조사 데이터란?

▶ 질병관리청에서 대한민국 국민의 건강 및 질병 실태에 대해서 약 20년 동안 전국 표본설문조사를 통해 조사하여 생성한 대한민국 국민의 보건/건강 빅데이터

국민건강영양조사 데이터의 사용과 활용 동의

국민건강영양조사 데이터를 사용하기 위해서는 질병관리청 사이트에서 **활용 동의 해야 함**

따라서, 본 경진대회에서는 **데이터의 직접 제공이 불가능**함

본 경진대회는 데이터셋을 생성할 수 있는 **데이터 프로세싱 코드를 대신 제공**함

경진대회 참가자 각자가 데이터 프로세싱 코드를 실행하여 데이터셋을 생성함

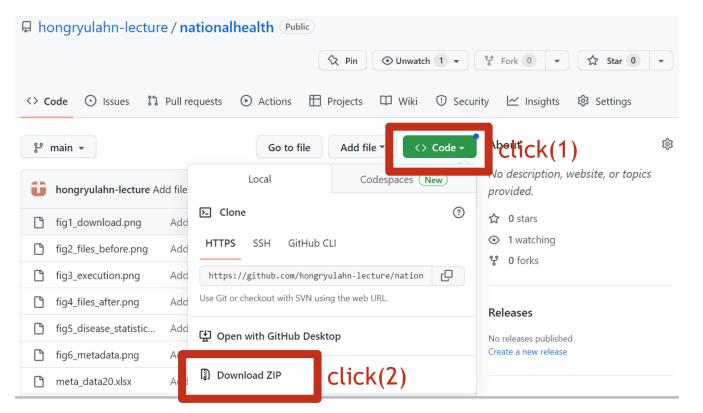
_1) 원본 데이터를 다운로드 받고, 로하는 데이터 프로세싱 코드를 실행성

2) 제공하는 데이터 프로세싱 코드를 실행하여 3) 경진대회 데이터셋을 생성하여 분석함

▶ 경진대회 데이터 생성

데이터 프로세싱 파일 다운로드

- ▶ 아래 링크로 접속하여 데이터 프로세싱 파일과 변수 설명 파일을 다운로드 받음
 - https://github.com/hongryulahn-lecture/nationalhealth



원본 데이터 다운로드

- ▶ 아래 링크로 접속하여 원본 데이터를 파일 12개를 다운로드 받음
 - https://knhanes.kdca.go.kr/knhanes/sub03/sub03_02_05.do
- ▶ 다운로드 받을 12개 파일은 2010년~2021년 기본DB 에 대한 sas7bat 파일임.
 - ► hn10_all.sas7bdat
 - hn11_all.sas7bdat
 - hn12_all.sas7bdat
 - hn13_all.sas7bdat
 - hn14_all.sas7bdat
 - hn15_all.sas7bdat
 - hn16_all.sas7bdat
 - hn17_all.sas7bdat
 - hn18_all.sas7bdat
 - hn19_all.sas7bdat
 - hn20_all.sas7bdat
 - hn21_all.sas7bdat



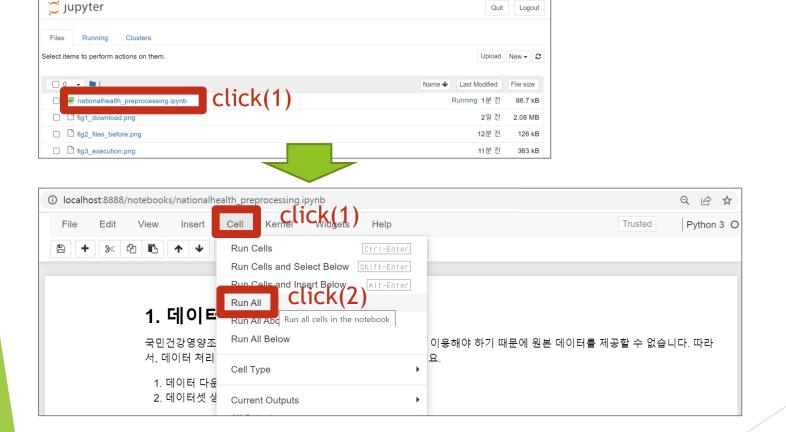
다운로드한 모든 파일을 한 폴더에 위치시킴

▶ 다운로드 받은 데이터프로세싱 파일과, 원본 데이터들을 한 폴더에 위치시킴

내 PC 》 로컬 디스크 (C:) 》 a			
이름	수정한 날짜	유형	크기
፴ fig1_download.png	2023-01-24 오후 8:47	알씨 PNG 파일	2,032KB
fig2_files_before.png	2023-01-27 오전 8:17	알씨 PNG 파일	123KB
fig3_execution.png	2023-01-27 오전 8:18	알씨 PNG 파일	355KB
fig4_files_after.png	2023-01-27 오전 8:17	알씨 PNG 파일	1,021KB
fig5_disease_statistics.png	2023-01-24 오후 3:51	알씨 PNG 파일	2,450KB
fig6_metadata.png	2023-01-27 오전 8:01	알씨 PNG 파일	1,903KB
hn10_all.sas7bdat	2022-06-10 오후 4:42	SAS Data Set	67,840KB
hn11_all.sas7bdat	2022-06-10 오후 4:43	SAS Data Set	63,504KB
hn12_all.sas7bdat	2022-06-10 오후 4:43	SAS Data Set	55,440KB
hn13_all.sas7bdat	2022-06-10 오후 4:44	SAS Data Set	67,788KB
hn14_all.sas7bdat	2022-08-05 오전 10:47	SAS Data Set	66,024KB
hn15_all.sas7bdat	2022-06-10 오후 4:44	SAS Data Set	69,300KB
hn16_all.sas7bdat	2023-01-13 오후 2:46	SAS Data Set	95,232KB
hn17_all.sas7bdat	2023-01-13 오후 12:58	SAS Data Set	130,560KB
hn18_all.sas7bdat	2023-01-13 오후 2:12	SAS Data Set	87,948KB
hn19_all.sas7bdat	2023-01-17 오전 6:52	SAS Data Set	93,492KB
hn20_all.sas7bdat	2023-01-17 오전 7:47	SAS Data Set	75,776KB
hn21_all.sas7bdat	2023-01-16 오후 1:37	SAS Data Set	69,300KB
meta_data20.xlsx	2023-01-11 오후 12:26	Microsoft Excel 워크	67KB
ationalhealth_preprocessing.ipynb	2023-01-27 오전 8:24	Jupyter 원본 파일	85KB

데이터 프로세싱 파일 실행

▶ 주피터 노트북에서 nationalhealth_preprocessing.ipynb을 열고 실행함.



데이터셋 생성 확인

▶ 정상 실행되었으면 nationalhealth_2010to2021.csv 파일이 생성됨

ㅐ PC ≯ 로컬 디스크 (C:) ≯ a			
이름	수정한 날짜	유형	크기
ipynb_checkpoints	2023-01-27 오전 8:26	파일 폴더	
fig1_download.png	2023-01-24 오후 8:47	알씨 PNG 파일	2,032KE
fig2_files_before.png	2023-01-27 오전 8:17	알씨 PNG 파일	123KE
fig3_execution.png	2023-01-27 오전 8:18	알씨 PNG 파일	355KE
fig4_files_after.png	2023-01-27 오전 8:17	알씨 PNG 파일	1,021KE
fig5_disease_statistics.png	2023-01-24 오후 3:51	알씨 PNG 파일	2,450KE
fig6_metadata.png	2023-01-27 오전 8:01	알씨 PNG 파일	1,903KE
hn10_all.sas7bdat	2022-06-10 오후 4:42	SAS Data Set	67,840KE
hn 11_all.sas 7bdat	2022-06-10 오후 4:43	SAS Data Set	63,504KE
hn 12_all.sas 7bdat	2022-06-10 오후 4:43	SAS Data Set	55,440KE
hn 13_all.sas 7bdat	2022-06-10 오후 4:44	SAS Data Set	67,788KE
hn14_all.sas7bdat	2022-08-05 오전 10:47	SAS Data Set	66,024KE
hn15_all.sas7bdat	2022-06-10 오후 4:44	SAS Data Set	69,300KE
hn16_all.sas7bdat	2023-01-13 오후 2:46	SAS Data Set	95,232KE
hn 17_all.sas 7bdat	2023-01-13 오후 12:58	SAS Data Set	130,560KE
hn 18_all.sas 7bdat	2023-01-13 오후 2:12	SAS Data Set	87,948KE
hn19_all.sas7bdat	2023-01-17 오전 6:52	SAS Data Set	93,492KE
hn20_all.sas7bdat	2023-01-17 오전 7:47	SAS Data Set	75,776KE
hn21_all.sas7bdat	2023-01-16 오후 1:37	SAS Data Set	69,300KE
meta data20.xlsx	2023-01-11 오후 12:26	Microsoft Excel 워크	67KE
nationalhealth_2010to2021.csv	2023-01-27 오전 8:26	Microsoft Excel 쉼표	28, 1 42KE
□ nationalnealtin_preprocessing.ipynb	2023-01-27 오전 8:26	Jupyter 원본 파일	85KE

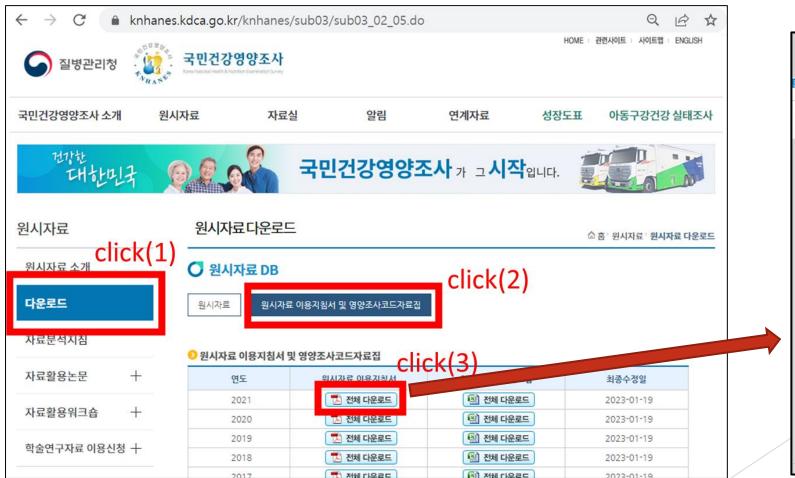
▶ 경진대회 데이터 설명

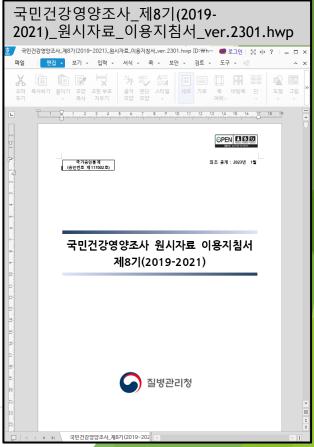
경진대회 데이터 파일

- nationalhealth_2010to2021.csv
 - ▶ 데이터셋 파일
 - ▶ github에서 프로세싱 코드를 실행하여 생성
- metadata_2020.xlsx
 - ▶ 변수 설명 파일(요약버전)
 - ▶ github에서 프로세싱 코드와 함께 제공
- ▶ 국민건강영양조사_제8기(2019-2021)_원시자료_이용지침서_ver.2301.hwp
 - ▶ 변수 설명 파일(원본버전)
 - ▶ 다음 슬라이드 참고하여 다운로드 받을 것

변수 설명 파일 받는 법

https://knhanes.kdca.go.kr/knhanes/sub03/sub03_02_05.do 에 접속하여 설명 파일 다운로드





데이터셋(nationalhealth_2010to2022.csv) 파일

- ▶ 32368 X 122 크기의 이차원 행렬 구조
- ▶ 행: 조사 대상자 : 32368명 (2011년~2020년까지 10년 동안의 조사 대상자)
- ▶ 열: 대상자에 대한 신상 및 건강 관련 변수
 - ▶ 영어 ID로 된 109개의 일반 변수와 한글 ID로 된 13개의 질병 변수로 구성됨

데이터셋(nationalhealth_2010to2022.csv) 파일

122개의 변수

109개의 일반 변수(영어ID)

13개의 질병 변수(한글ID)

	Α	В	С	D	E	F
1	ID	ID_fam	year	region	town_t	sex
2	b'A308780	b'A308780	2010	1	1	1
3	b'A309099	b'A309099	2010	1	1	2
4	b'A309460	b'A309460	2010	1	1	2
5	b'A309460	b'A309460	2010	1	1	1
6	b'A310439	b'A310439	2010	1	1	2
7	b'A310980	b'A310980	2010	1	1	1
8	b'A310980	b'A310980	2010	1	1	2
9	b'A313020	b'A313020	2010	1	1	1
10	b'A313080	b'A313080	2010	1	1	2
11	b'A313240	b'A313240	2010	1	1	1
12	b'A313240	b'A313240	2010	1	1	2
13	b'A314539	b'A314539	2010	1	1	1
35620	b'R904239	b'R904239	2021	8	2	1
35621	b'R904239	b'R904239	2021	8	2	2
35622	b'R904303	b'R904303	2021	8	2	2
35623	b'R904310	b'R904310	2021	8	2	1
35624	b'R904322	b'R904322	2021	8	2	2
35625	b'R904322	b'R904322	2021	8	2	2
35626	b'R904322	b'R904322	2021	8	2	1
35627	b'R904332	b'R904332	2021	8	2	1
35628	b'R904346	b'R904346	2021	8	2	2
35629	b'R904353	b'R904353	2021	8	2	1
35630	b'R904353	b'R904353	2021	8	2	2

•	DB	DC	DD	DE	DF	DG	DH	DI
	N_B1	N_B2	N_NIAC	N_VITC	비만	고혈압	당뇨병	고콜레스
	1.795274	1.067851	29.15508	165.4297	1	0	1	0
	0.69894	0.314588	19.9291	49.12153	0	1	1	0
	0.446677	0.269494	6.259597	39.23053	1	1	1	0
	0.493381	0.591819	11.20549	46.50654	1	0	1	0
	2.136759	2.144746	20.181	339.5929	0	1	1	0
	0.604341	1.055412	7.415678	19.93856	0	0	1	0
	0.629514	0.809784	6.255108	29.1494	0	1	1	0
	0.980365	0.472175	7.970716	105.8412	0	0	1	0
	1.473793	0.807193	11.24571	131.243	1	1	1	0
	1.099274	1.064746	17.2626	90.14383	1	1	1	0
	1.870165	1.482319	16.48341	70.66173	1	1	1	0
	0.25228	0.145264	3.591637	11.16345	0	0	1	0
	1.618756	1.589698	19.88959	109.6958	1	1	0	1
	0.623761	0.440142	4.259609	94.09098	0	1	0	0
	0.832882	1.421409	10.55521	17.17904	0	0	0	1
	0.566141	1.081714	7.410071	17.62967	0	0	0	0
	0.940095	0.978504	10.08709	29.29815	0	1	1	1
	1.155178	0.891093	8.977854	43.02104	1	0	0	0
	1.714108	1.215324	15.84191	112.9737	1	1	0	0
	0.338357	0.608196	3.326717	27.34886	0	0	0	0
	0.769159	1.251299	10.76729	106.7738	0	0	0	0
	0.881561	1.532354	14.06587	21.53251	1	1	0	0
	0.986966	1.14764	10.98538	39.2254	0	0	0	0

DO	DP	DQ	DR
천식	아토피피	골관절염	우울증
0	-1	-1	-1
-1	-1	-1	-1
1	-1	1	-1
-1	-1	-1	1
-1	-1	1	-1
-1	-1	1	-1
-1	-1	1	-1
-1	-1	-1	-1
1	-1	-1	-1
-1	-1	-1	-1
-1	-1	-1	-1
-1	-1	-1	-1
0	0	0	0
0	0	0	0
0	0	0	0
0	0	0	0
0	0	1	0
0	0	0	0
0	0	0	0
0	0	0	0
0	0	0	0
0	0	0	0
0	0	0	0

35,629명의 대상자 (2010년~2021년)

데이터셋(nationalhealth_2010to2021.csv) 그 일반 변수

- ▶ 영어 변수 ID인 109개의 변수
- ▶ 조사 대상자의 성별, 나이, 혈압, 탄수화물 섭취량 등의 신체, 건강, 영양에 대한 조사 및 측<mark>정한 변수</mark>
- ▶ 각 ID와 값의 의미에 대해서는 'metadata.xlsx'(요약본) 파일과 '국민건강영양조사_제8기(2019-2021)_원시자료_이용지침서_ver.2301.hwp'(상세본) 파일을 참고할 것

데이터셋(nationalhealth_2010to2021.csv) 부음 - 일반 변수

예) educ (교육수준) 변수에 대해서 요약본 파일과 상세본 파일에서 정보를 확인할 수 있음.

metadata.xlsx(요약본) 파일

1 variat ▼	variable description	*	option description
266 educ	교육수준: 학력		1:서당/한학 2:무학 3:초등학교 4:중학교 5:고등학교 6:2년/3년제 대학 7:4년제 대학 8:대학원

▶ 국민건강영양조사_제8기(2019-2021)_원시자료_이용지침서_ver.2301.hwp(상세본) 파일

문항번호	변수유형	변수명	변수설명	내 용
1	N(2)	educ	교육수준: 학립	1. 서당/한학 2. 부학 3. 초등학교 4. 중학교 5. 고등학교 6. 2년/3년제 대학 7. 4년제 대학 8. 대학원 88. 비해당(소아) 99. 모름, 부용답

설문 및 변수			
여 성인, 청소년			
1. 귀하께서는 학교를 어디까	지 다니셨습니까? 혹은 다니고 계	[십니까? (educ)	
서당/한학	2 부학	3 초등학교	4 중학교
5 고통학교	6 2년/3년제 대학	7 4년제 대학	8 대학원
의 파유치파	তা হতাও্ত্ৰ পাল	II 45√1 41∓	ण भाषाच

데이터셋(nationalhealth_2010to2021.csv) 꾸

- 질병 변수

- ▶ 한글 ID로 된 13개의 변수
- 질병 유무에 대한 변수(질병 유무 결정의 기준은 국건영 통계자료를 참고함)
 - 1) 비만
 - ₂₎ 고혈압
 - 3) 당뇨병
 - 4) 고콜레스테롤혈증
 - 5) 고중성지방혈증
 - 6) B형간염
 - 7) 빈혈
 - 8) 뇌졸중
 - 9) 협심증또는심근경색증
 - 10) 전식
 - 11) 아토피피부염
 - 12) 골관절염
 - 13) 우울증



데이터셋(nationalhealth_2010to2021.csv) 부음 - 결측값

- ▶ 양수는 보통 값을 의미
- ▶ -1은 '해당없음'을 의미
 - ▶ 예: '출산 여부' 변수에 대해서, 출산한 여성은 1, 미출산 여성은 0, 남자는 -1(해당없음) 값을 가짐.
- ▶ -2는 '응답 안함 또는 모름'을 의미
 - ▶ 예: '교육 수준' 변수에 대해서, 응답 안한 사람은 -2값을 가짐.

경진대회 분석 주제

- ► 제공하는 국민건강영양조사 데이터에 대해서 '특정 질병에 대한 중요 요인'
 - '환자 클러스터링'
 - '질병 클러스터링'
 - '성별/나이에 따른 신체/질병 관련 특성'
 - '연도 변화에 따른 신체/질병 특성의 변화' 등 자유롭게 주제를 정해 분석하여 결과를 제출 (언급한 것 외 주제도 모두 가능)

▶ 경진대회 세부 규칙

제출물 및 제출 방법

▶ 제출물

- 1) 데이터 분석 결과에 대한 발표자료(pptx 또는 pdf 파일)
 - ▶ 15~20분 발표 정도 분량
 - ▶ 제목 페이지에 팀 이름, 팀 구성원의 이름과 학번 정보 포함할 것
- 2) 분석 코드(ipynb 또는 r 파일)

▶ 제출 방법

- ▶ 분석 결과 제출 기한 (2023년 2/17일 23:59) 까지
- ▶ hongryulahn@gmail.com(안홍렬 교수)으로 이메일 제출

시상내역

- ▶ 상금
 - ▶ 결과발표: 2023년 2/22일
 - ▶ 상금은 "국민건강영양조사 데이터 분석" 주제에 대해서 순위 선정하여 지급
 - ▶ 1등(1팀) 30만원
 - ▶ 2등(2팀) 각 15만원
 - ▶ 3등(3팀) 각 5만원
 - ▶ 참가팀의 결과물 품질에 따라 총 상금 내에서 상금 배분은 조정될 수 있음
- ▶ 경진대회 최종 순위 및 시상식
 - ▶ 경진대회 최종 순위는 2022년 12월 종료된 자연어분석 공모전 시상자와 경합하여 선정함 (즉, 국민건강영양조사 데이터분석 상금 지급 순위는 경진대회 최종 순위가 아님)
 - ▶ 경진대회 최종 순위 선정자에 대해서 3월 중 상장을 시상함

경진대회 및 데이터 관련 문의

- ▶ 전체 문의
 - ▶ 경진대회 참가하는 사람은 아래 경진대회 오픈 단체 카톡방에 참가할 것.
 - https://open.kakao.com/o/g8gesS0e (참여코드:data)
 - ▶ 경진대회 관련 문의는 해당 단톡방에 질문할 것
 - ▶ 추가적 공지사항은 해당 단체 카톡방에 공지 예정
- ▶ 개인 문의
 - ▶ 개인적 문의가 필요한 경우는 안홍렬 교수의
 - ▶ 개인 오픈 카카오톡 채팅(https://open.kakao.com/o/sr6rod3b)이나
 - ▶ 이메일(<u>hongryulahn@gmail.com</u>)로 문의