Big Data Project
Topic Proposal

11조

강유민, 김근하, 김대원

Topic

개인의 **기본정보**와 **건강정보**를 통해 분류된
특정집단의 **건강상태를 분석**하고,

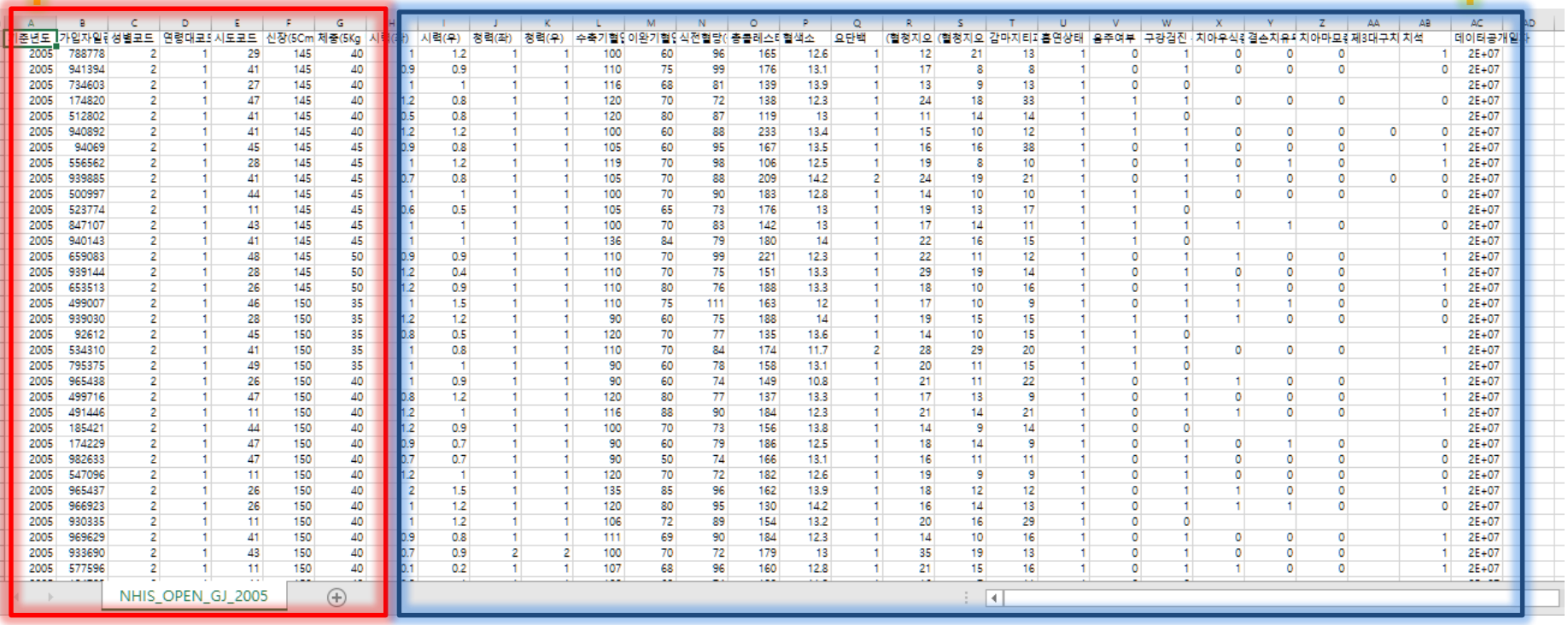
특정개인의 기본정보를 유사집단과 비교하여
특정개인의 **건강상태를 예측**

Data

- **국민 건강 보험 공단**에서 제공한 데이터
- **건강검진 정보** & **진료내역 정보** 데이터
- **연도별 데이터** 제공 (2005년 ~ 2015년)
- 진료 및 건강검진을 수진한 국민건강보험 가입자 100만명 **무작위 추출**로 표본 형성

Data – 건강 검진 정보 데이터

29 columns



기준년도	가입자일련	성별코드	연령대코드	시도코드	신장(5Cm	체중(5Kg	시력(우)	청력(우)	청력(좌)	수축기혈	이완기혈	식전혈당(	총콜레스테	혈색소	요단백	(혈청지오	(혈청지오	감마지티	흡연상태	음주여부	구강검진	치아우식	결손치유	치아마모	제3대구치	치석	데이터공개일
2005	788778	2	1	29	145	40	1	1.2	1	1	100	60	96	165	12.6	1	12	21	13	1	0	1	0	0	0	1	2E+07
2005	941394	2	1	41	145	40	0.9	0.9	1	1	110	75	99	176	13.1	1	17	8	8	1	0	1	0	0	0	0	2E+07
2005	734603	2	1	27	145	40	1	1	1	1	116	68	81	139	13.9	1	13	9	13	1	0	0					2E+07
2005	174820	2	1	47	145	40	1.2	0.8	1	1	120	70	72	138	12.3	1	24	18	33	1	1	1	0	0	0	0	2E+07
2005	512802	2	1	41	145	40	0.5	0.8	1	1	120	80	87	119	13	1	11	14	14	1	1	0					2E+07
2005	940892	2	1	41	145	40	1.2	1.2	1	1	100	60	88	233	13.4	1	15	10	12	1	1	1	0	0	0	0	2E+07
2005	94069	2	1	45	145	45	0.9	0.8	1	1	105	60	95	167	13.5	1	16	16	38	1	0	1	0	0	0	1	2E+07
2005	556562	2	1	28	145	45	1	1.2	1	1	119	70	98	106	12.5	1	19	8	10	1	0	1	0	1	0	1	2E+07
2005	939885	2	1	41	145	45	0.7	0.8	1	1	105	70	88	209	14.2	2	24	19	21	1	0	1	1	0	0	0	2E+07
2005	500997	2	1	44	145	45	1	1	1	1	100	70	90	183	12.8	1	14	10	10	1	1	1	0	0	0	0	2E+07
2005	523774	2	1	11	145	45	0.6	0.5	1	1	105	65	73	176	13	1	19	13	17	1	1	0					2E+07
2005	847107	2	1	43	145	45	1	1	1	1	100	70	83	142	13	1	17	14	11	1	1	1	1	1	0	0	2E+07
2005	940143	2	1	41	145	45	1	1	1	1	136	84	79	180	14	1	22	16	15	1	1	0					2E+07
2005	659083	2	1	48	145	50	0.9	0.9	1	1	110	70	99	221	12.3	1	22	11	12	1	0	1	1	0	0	1	2E+07
2005	939144	2	1	28	145	50	1.2	0.4	1	1	110	70	75	151	13.3	1	29	19	14	1	0	1	0	0	0	1	2E+07
2005	653513	2	1	26	145	50	1.2	0.9	1	1	110	80	76	188	13.3	1	18	10	16	1	0	1	1	0	0	1	2E+07
2005	499007	2	1	46	150	35	1	1.5	1	1	110	75	111	163	12	1	17	10	9	1	0	1	1	1	0	0	2E+07
2005	939030	2	1	28	150	35	1.2	1.2	1	1	90	60	75	188	14	1	19	15	15	1	1	1	1	0	0	0	2E+07
2005	92612	2	1	45	150	35	0.8	0.5	1	1	120	70	77	135	13.6	1	14	10	15	1	1	0					2E+07
2005	534310	2	1	41	150	35	1	0.8	1	1	110	70	84	174	11.7	2	28	29	20	1	1	1	0	0	0	1	2E+07
2005	795375	2	1	49	150	35	1	1	1	1	90	60	78	158	13.1	1	20	11	15	1	1	0					2E+07
2005	965438	2	1	26	150	40	1	0.9	1	1	90	60	74	149	10.8	1	21	11	22	1	0	1	1	0	0	1	2E+07
2005	499716	2	1	47	150	40	0.8	1.2	1	1	120	80	77	137	13.3	1	17	13	9	1	0	1	0	0	0	1	2E+07
2005	491446	2	1	11	150	40	1.2	1	1	1	116	88	90	184	12.3	1	21	14	21	1	0	1	1	0	0	1	2E+07
2005	185421	2	1	44	150	40	1.2	0.9	1	1	100	70	73	156	13.8	1	14	9	14	1	0	0					2E+07
2005	174229	2	1	47	150	40	0.9	0.7	1	1	90	60	79	186	12.5	1	18	14	9	1	0	1	0	1	0	0	2E+07
2005	982633	2	1	47	150	40	0.7	0.7	1	1	90	50	74	166	13.1	1	16	11	11	1	0	1	0	0	0	0	2E+07
2005	547096	2	1	11	150	40	1.2	1	1	1	120	70	72	182	12.6	1	19	9	9	1	0	1	0	0	0	0	2E+07
2005	965437	2	1	26	150	40	2	1.5	1	1	135	85	96	162	13.9	1	18	12	12	1	0	1	1	0	0	1	2E+07
2005	966923	2	1	26	150	40	1	1.2	1	1	120	80	95	130	14.2	1	16	14	13	1	0	1	1	1	0	0	2E+07
2005	930335	2	1	11	150	40	1	1.2	1	1	106	72	89	154	13.2	1	20	16	29	1	0	0					2E+07
2005	969629	2	1	41	150	40	0.9	0.8	1	1	111	69	90	184	12.3	1	14	10	16	1	0	1	0	0	0	1	2E+07
2005	933690	2	1	43	150	40	0.7	0.9	2	2	100	70	72	179	13	1	35	19	13	1	0	1	0	0	0	1	2E+07
2005	577596	2	1	11	150	40	0.1	0.2	1	1	107	68	96	160	12.8	1	21	15	16	1	0	1	1	0	0	1	2E+07

NHIS_OPEN_GJ_2005

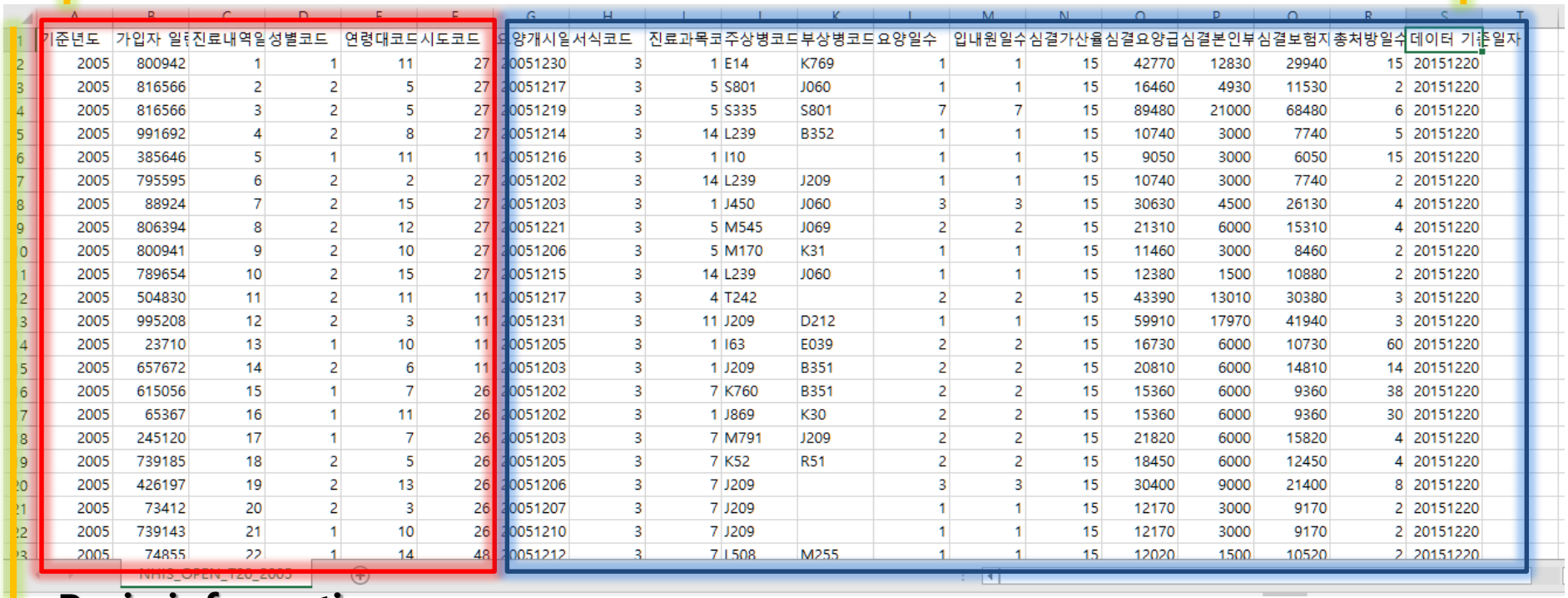
Basic information
- 성별, 연령, 시·도, 신장, 체중

More than 1million rows

Examination information
- 시력, 청력, 혈압, 혈당, 콜레스테롤, 요단백 … …

Data – 진료 내역 정보 데이터

19 columns



기준년도	가입자 일련	진료내역일	성별코드	연령대코드	시도코드	양개시일	서식코드	진료과목코	주상병코드	부상병코드	요양일수	입내원일수	심결가산율	심결요양급	심결본인부	심결보험지	총처방일수	데이터 기준일자
2005	800942	1	1	11	27	20051230	3	1	E14	K769	1	1	15	42770	12830	29940	15	20151220
2005	816566	2	2	5	27	20051217	3	5	S801	J060	1	1	15	16460	4930	11530	2	20151220
2005	816566	3	2	5	27	20051219	3	5	S335	S801	7	7	15	89480	21000	68480	6	20151220
2005	991692	4	2	8	27	20051214	3	14	L239	B352	1	1	15	10740	3000	7740	5	20151220
2005	385646	5	1	11	11	20051216	3	1	I10		1	1	15	9050	3000	6050	15	20151220
2005	795595	6	2	2	27	20051202	3	14	L239	J209	1	1	15	10740	3000	7740	2	20151220
2005	88924	7	2	15	27	20051203	3	1	J450	J060	3	3	15	30630	4500	26130	4	20151220
2005	806394	8	2	12	27	20051221	3	5	M545	J069	2	2	15	21310	6000	15310	4	20151220
2005	800941	9	2	10	27	20051206	3	5	M170	K31	1	1	15	11460	3000	8460	2	20151220
2005	789654	10	2	15	27	20051215	3	14	L239	J060	1	1	15	12380	1500	10880	2	20151220
2005	504830	11	2	11	11	20051217	3	4	T242		2	2	15	43390	13010	30380	3	20151220
2005	995208	12	2	3	11	20051231	3	11	J209	D212	1	1	15	59910	17970	41940	3	20151220
2005	23710	13	1	10	11	20051205	3	1	I63	E039	2	2	15	16730	6000	10730	60	20151220
2005	657672	14	2	6	11	20051203	3	1	J209	B351	2	2	15	20810	6000	14810	14	20151220
2005	615056	15	1	7	26	20051202	3	7	K760	B351	2	2	15	15360	6000	9360	38	20151220
2005	65367	16	1	11	26	20051202	3	1	J869	K30	2	2	15	15360	6000	9360	30	20151220
2005	245120	17	1	7	26	20051203	3	7	M791	J209	2	2	15	21820	6000	15820	4	20151220
2005	739185	18	2	5	26	20051205	3	7	K52	R51	2	2	15	18450	6000	12450	4	20151220
2005	426197	19	2	13	26	20051206	3	7	J209		3	3	15	30400	9000	21400	8	20151220
2005	73412	20	2	3	26	20051207	3	7	J209		1	1	15	12170	3000	9170	2	20151220
2005	739143	21	1	10	26	20051210	3	7	J209		1	1	15	12170	3000	9170	2	20151220
2005	74855	22	1	14	48	20051212	3	7	L508	M255	1	1	15	12020	1500	10520	2	20151220

Basic information

- 진료내역 번호, 성별, 연령, 시·도

Treatment information

- 요양 개시일, 진료과목, 주상병코드, 부상병코드, 요양일수, 입 내원 일수
- … …

More than 1million rows

Data – considerations

- 진료 내역 데이터의 경우 해당 연도에 요양(병/의원)기관 기록이 1건 이상 있는 가입자 100만명을 표본으로 선정
- 진료 내역 데이터 셋의 크기를 축소하기 위해 구간 분포 비율을 유지한 채로 정제 과정 거침

Analysis – expectation

1st. Data purification
- 기본 분류군 형성(Ex. 성별, 연령, 흡연여부)

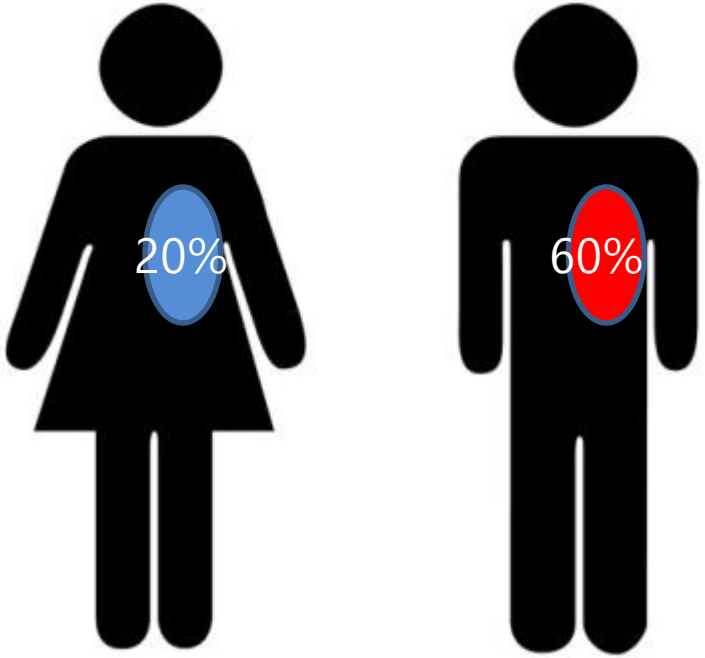
2nd. Exploratory Data Analysis
- 상관관계 도출

(Ex. 30대 성별에 따른 폐암 발병률, 연도별 당뇨환자 추이)

3rd . Visualization
- 기본 분류군에 따른 건강상태
- 특정 개인의 정보와 유사 집단 비교를 통한 건강 상태 예측

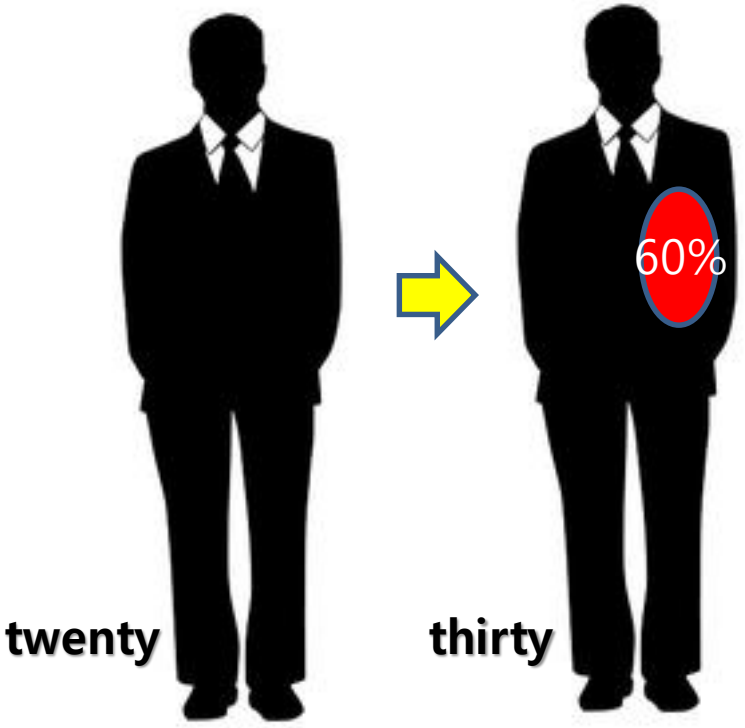
Result – expectation

Smoker, Thirty



Result 1

Smoker, 180cm, 70kg, Daegu



Result 2

Thank you