- 빅데이터 프로젝트 최종 보고서 -

팀: 11조
201511006: 강유민
201511021: 김근하
201511024: 김대원

Ⅰ. 프로젝트 주제 및 목표

ⅰ. 초기 주제 및 목표 (프로젝트 제안 당시)

개인의 기본 정보와 건강 정보를 통해 분류된 특정집단의 건강상태를 분석하고,
특정 개인의 기본정보를 유사집단과 비교하여 특정개인의 건강상태를 예측한다.

ⅱ. 최종 주제 및 목표 (프로젝트 최종 발표)

진료내역 데이터를 이용하여 질병 유형 및 발병 경향을 분석하고,
이를 통해 의료 산업의 마케팅 전략에 도움이 될 수 있는 정보를 제공한다.

Ⅱ. 데이터

ⅰ. 건강 검진 데이터 (2005 ~ 2015)

1. 데이터 생성 방식 및 범위

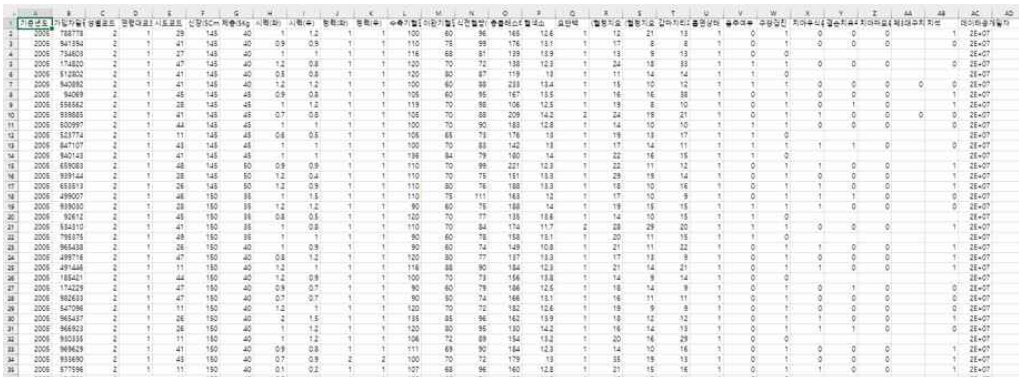
해당 년도에 건강검진을 수진한 국민건강보험가입자 100만 명을 무작위로 선별하고,
항목 선정 과정을 거쳐 선정된 가입자의 기본정보와 검진결과정보를 추출하여 국민건강정보 데이터셋을 구성하였다.

2. 개념

2002년부터 2015년까지, 국민건강보험의 직장가입자와 40세 이상의 피부양자, 세대주인 지역가입자와 40세 이상의 지역가입자의 「일반건강검진」 결과와 이들 일반건강검진 대상자 중에 만40세와 만66세에 도달한 이들이 받게 되는「생애전환기건강진단」의 결과이다. 직장가입자 및 지역세대주는 연령제한 없이 건강검진 대상자가 되며 본 건강검진정보에는 20세 미만의 검진결과는 제외하고 구축하였다.

3. 구성

총 34개의 변수로 가입자 일련번호와 ① **수진자 기본정보** : 성, 연령, 거주지 시도코드,
② **건강검진결과 및 문진정보** : 신체, 몸무게, 허리둘레 등 신체사이즈 정보와 혈압, 혈당, 콜레스테롤, 요단백, 감마지피티와 같은 병리검사결과 시력과 청력, 구강검사와 같은 진단검사결과 그 외 음주와 흡연 여부에 대한 문진결과로 구성되어있다.



ⅱ. 진료 내역 데이터 (2013 ~ 2015)

1. 데이터 생성 방식 및 범위

국민건강보험가입자 중 해당 년도에 요양(병/의원)기관으로 부터의 진료내역이 1건 이상 있는 가입자 100만 명을 무작위로 선별하고, 항목 선정 과정을 거쳐 선정된 해당 가입자의 기본정보와 진료정보를 추출하여 1차 데이터셋을 구성하였다. 1차 데이터셋의 크기를 축소하기 위하여 구간분포비율을 최대한 유지한 상태에서 데이터 정제 작업을 거쳐 최종적인 진료내역정보 데이터셋을 구성하였다.

2. 개념

2002년부터 2015년까지의 각 연도별 수진자 100만 명에 대한 기본정보(성, 연령대, 시도코드 등)와 진료내역(진료과목코드, 주상병 코드, 요양일수, 총처방일수 등)으로 구성된 개방 데이터이다.

3. 구성

총 19개의 변수로 가입자 일련번호와 진료내역 일련번호, ① **수진자 기본정보** : 성, 연령, 거주지 시도코드, ② **진료상세 정보** : 주상병, 부상병, 요양일수, 입·내원일수, 총 처방일수 그리고 ③ **요양급여 청구 심사 결과**에 대한 정보로 구성되어 있다.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1	기준년도	가입자 일	진료내역일	성별코드	연령대코드	시도코드	요양개시일	서식코드	진료과목코	주상병코드	부상병코드	요양일수	입내원일수	심결가산율	심결요양급	심결본인부	심결보험자	총처방일수	데이터 기준일자	
2	2005	800942	1	1	11	27	20051230	3	1	E14	K769	1	1	15	42770	12830	29940	15	20151220	
3	2005	816566	2	2	5	27	20051217	3	5	S801	J060	1	1	15	16460	4930	11530	2	20151220	
4	2005	816566	3	2	5	27	20051219	3	5	S335	S801	7	7	15	89480	21000	68480	6	20151220	
5	2005	991692	4	2	8	27	20051214	3	14	L239	B352	1	1	15	10740	3000	7740	5	20151220	
6	2005	385646	5	1	11	11	20051216	3	1	I10		1	1	15	9050	3000	6050	15	20151220	
7	2005	795595	6	2	2	27	20051202	3	14	L239	J209	1	1	15	10740	3000	7740	2	20151220	
8	2005	88924	7	2	15	27	20051203	3	1	J450	J060	3	3	15	30630	4500	26130	4	20151220	
9	2005	806394	8	2	12	27	20051221	3	5	M545	J069	2	2	15	21310	6000	15310	4	20151220	
10	2005	800941	9	2	10	27	20051206	3	5	M170	K31	1	1	15	11460	3000	8460	2	20151220	
11	2005	789654	10	2	15	27	20051215	3	14	L239	J060	1	1	15	12380	1500	10880	2	20151220	
12	2005	504830	11	2	11	11	20051217	3	4	T242		2	2	15	43390	13010	30380	3	20151220	
13	2005	995208	12	2	3	11	20051231	3	11	J209	D212	1	1	15	59910	17970	41940	3	20151220	
14	2005	23710	13	1	10	11	20051205	3	1	I63	E039	2	2	15	16730	6000	10730	60	20151220	
15	2005	657672	14	2	6	11	20051203	3	1	J209	B351	2	2	15	20810	6000	14810	14	20151220	
16	2005	615056	15	1	7	26	20051202	3	7	K760	B351	2	2	15	15360	6000	9360	38	20151220	
17	2005	65367	16	1	11	26	20051202	3	1	J869	K30	2	2	15	15360	6000	9360	30	20151220	
18	2005	245120	17	1	7	26	20051203	3	7	M791	J209	2	2	15	21820	6000	15820	4	20151220	
19	2005	739185	18	2	5	26	20051205	3	7	K52	R51	2	2	15	18450	6000	12450	4	20151220	
20	2005	426197	19	2	13	26	20051206	3	7	J209		3	3	15	30400	9000	21400	8	20151220	
21	2005	73412	20	2	3	26	20051207	3	7	J209		1	1	15	12170	3000	9170	2	20151220	
22	2005	739143	21	1	10	26	20051210	3	7	J209		1	1	15	12170	3000	9170	2	20151220	
23	2005	74855	22	1	14	48	20051212	3	7	L508	M255	1	1	15	12020	1500	10520	2	20151220	

NHIS_OPEN_T20_2005

ⅲ. 시·도별 인구수 데이터 (2013 ~ 2015)

통계청에서 제공한 자료로서 2013년부터 2015년까지의 각 연도에 따른 시·도별 인구수를 제공하고 있다. 데이터 활용을 위하여 2013년부터 2015년까지의 각 시·도별 인구수를 합산하여 새롭게 가공하였다.

[통계청 제공 데이터]

	A	B	C	D	E	F	G	H	I
1	통계표명: 지역별 인구 및 인구밀도								
2	단위: 천명, 명/㎢								
3		2012		2013		2014		2015	
4		인구	인구밀도	인구	인구밀도	인구	인구밀도	인구	인구밀도
5	계	50,200	501	50,429	503	50,747	506	51,015	509
6	서울	10,036	16,583	9,990	16,507	9,975	16,482	9,941	16,425
7	부산	3,462	4,498	3,456	4,489	3,452	4,485	3,452	4,484
8	대구	2,480	2,807	2,476	2,802	2,475	2,801	2,469	2,794
9	인천	2,794	2,684	2,830	2,718	2,862	2,732	2,883	2,748
10	광주	1,504	3,000	1,504	3,000	1,505	3,002	1,506	3,005
11	대전	1,540	2,852	1,545	2,860	1,553	2,879	1,542	2,860
12	울산	1,125	1,061	1,137	1,073	1,151	1,085	1,164	1,097
13	세종	102	220	118	255	132	285	187	403
14	경기	11,974	1,177	12,126	1,192	12,282	1,207	12,423	1,221
15	강원	1,504	90	1,506	89	1,510	90	1,517	90
16	충북	1,553	210	1,565	211	1,578	213	1,589	215
17	충남	2,043	249	2,062	251	2,088	254	2,103	256
18	전북	1,817	225	1,821	226	1,829	227	1,835	227
19	전남	1,782	145	1,784	145	1,792	146	1,797	146
20	경북	2,656	140	2,661	140	2,671	140	2,678	141
21	경남	3,265	310	3,278	311	3,307	314	3,330	316
22	제주	561	303	570	308	583	315	599	324
23	수도권	24,805	2,099	24,946	2,111	25,119	2,124	25,247	2,134
24	출처: 통계청 『장래인구추계 시도편 : 2015-2045』, 국토교통부 『지적통계』								
25	주석: * 수도권 : 서울, 인천, 경기								
26	* 통계청 『장래인구추계 시도편 : 2015-2045』의 시도별 인구와 국토교통부 『지적통계』의 시도별 국토면적을 기초로 작성								
27	* 시도별 장래인구추계는 2017년에 작성된 자료임								

[가공 데이터]

	A	B
1	city	number
2	11	29906000
3	26	10360000
4	27	7420000
5	28	8575000
6	29	4515000
7	30	4640000
8	31	3452000
9	41	36831000
10	42	4533000
11	43	4732000
12	44	6253000
13	45	5485000
14	46	5373000
15	47	8010000
16	48	9915000
17	49	1752000

ⅳ. 지역별 진료과목 데이터

대한 병원 협회에서 제공한 데이터로서 지역별 진료과목 현황을 제공하고 있다. 따로 파일 형식으로 제공된 형태가 아니어서 분석에 필요한 과목(내과, 소아청소년과, 이비인후과, 정형외과, 안과) 총 5가지 진료과목에 대해서만 추출하여 새롭게 가공하였다.

[가공 데이터]

	city	Internal m	pediatrics	otolaryngo	orthopedi	ophthalmology
2	11	175	104	54	141	46
3	26	132	77	36	98	16
4	27	99	58	21	66	12
5	28	48	29	12	36	6
6	29	46	30	10	29	7
7	30	50	17	12	32	8
8	31	41	23	12	23	4
9	41	240	134	67	168	3
10	42	47	23	19	37	11
11	43	46	21	8	29	7
12	44	62	27	16	39	8
13	45	74	27	15	51	8
14	46	82	33	22	55	13
15	47	100	33	26	55	16
16	48	134	76	50	89	14
17	49	8	6	4	7	4

Ⅲ. 분석 과정 및 세부 내용

ⅰ. 건강 검진 데이터 분석

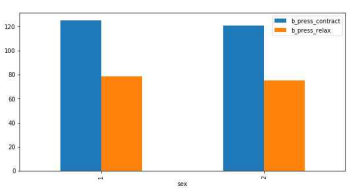
1. 수진자의 기본 정보(성, 연령, 거주지), 신체 정보(키, 몸무게), 음주여부와 흡연여부 각각을 변수로 하여 그에 따른 건강검진 결과 및 문진 정보(혈압, 혈당, 콜레스테롤, 요단백, 감마 지피티와 같은 병리검사결과 시력과 청력과 같은 진단검사결과)를 각각 살펴보았다.
2. 1.의 결과를 연도별로 비교하여 어떤 변화가 있는지 살펴보았다.
3. 수진자의 기본 정보(성, 연령, 거주지), 신체 정보(키, 몸무게), 음주여부와 흡연여부를 두 개씩 묶어서 변수로 지정하고 그에 따른 건강검진 결과 및 문진 정보를 살펴보았다.
4. 머신러닝을 통해 특정 년도의 수진자의 기본 정보(성, 연령, 거주지), 신체 정보(키, 몸무게), 음주여부와 흡연여부를 n 개씩 묶어서 변수로 지정했을 때, 그에 따른 건강검진 결과 및 문진 정보를 학습시킨 후에 그 다음 년도의 데이터와 비교해보았다.
5. 앞서 개요에서 설정한 변수로 분류했을 때, 해당 데이터의 수가 100만개의 1%미만(1만개)인 그룹은 데이터의 수가 충분하지 않다고 판단하여 분석해서 제외하였다. 따라서 연령대는 1(20~24세)~12(75~79세)까지, 키는 145cm~180cm(5cm 단위)까지, 몸무게는 45kg~90kg (5kg) 단위까지를 분석하였다.

◆ 1변수(기본 정보)에 대한 건강 상태 경향 분석

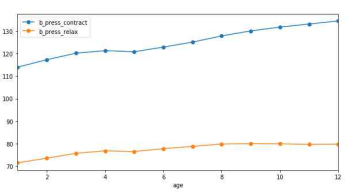
각 연도별 분석 결과 경향 추이가 모두 동일하게 나타나 2005년도 데이터를 기반으로 분석 결과를 설명하겠다. 각 항목별 분석 중 상관관계를 보이는 항목에 대해서만 서술하겠다.

Ⓐ **혈압 분석**

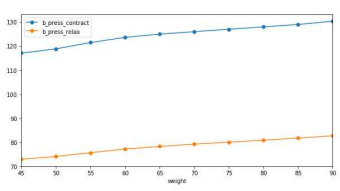
성별 vs 혈압



나이 vs 혈압



몸무게 vs 혈압



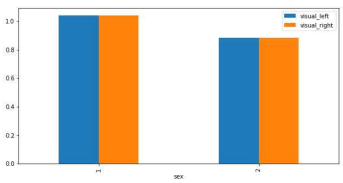
수축기 혈압은 심장이 수축할 때 현관에 가해지는 압력으로서 고혈압 진단에 사용된다. 이때 수축기 혈압이 140mmHg 이상인 경우 고혈압, 120mmHg 이상인 경우 고혈압 전 단계를 의심한다.

이완기 혈압은 심장이 이완할 때 혈관에 가해지는 압력으로서 고혈압 진단에 사용된다. 이완기 혈압이 90mmHg 이상인 경우 고혈압, 80mmHg 이상인 경우 고혈압 전 단계를 의심한다.

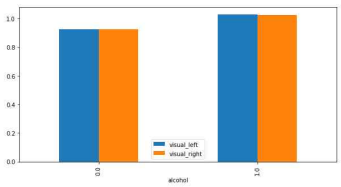
분석 결과 성별, 나이, 몸무게에서 혈압과의 상관관계를 보였다. 먼저 성별의 경우 남성(125.25, 78.81)이 평균적으로 여성(120.53, 75.00)에 비해 혈압(수축기, 이완기)이 높게 측정 되었다. 몸무게의 경우 몸무게가 증가함에 따라 혈압이 점차 증가하는 추세를 보였다. 특히, 75kg 이상(126.99, 80.04)에서 수축기, 이완기 혈압 모두 고혈압 전단계로 의심되는 수치를 보였다. 나이의 경우 연령이 증가함에 따라 혈압이 증가하는 추세를 보였다. 특히, 50세 이상부터 증가 추세 기울기가 증가하였고 60세 이상(130.10 80.00)부터는 수축기, 이완기 혈압 모두 고혈압 전단계로 의심되는 수치를 나타내었다.

Ⓑ 시력 분석

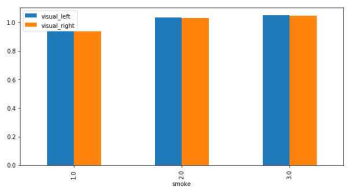
성별 vs 시력



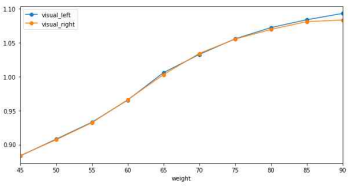
음주여부 vs 시력



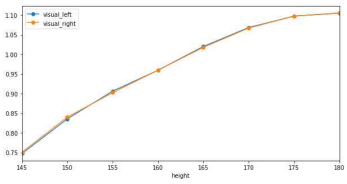
흡연여부 vs 시력



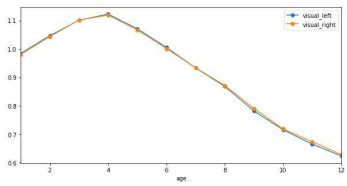
몸무게 vs 시력



키 vs 시력



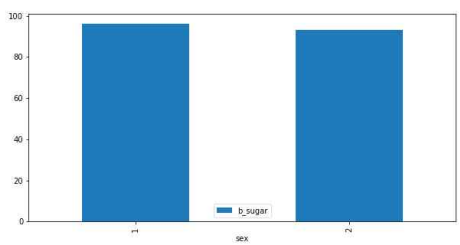
나이 vs 시력



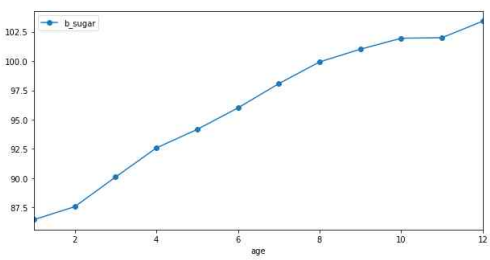
분석 결과 도시를 제외한 여섯 가지 기본 요소에 대하여 모두 상관관계를 보였다. 먼저 성별 분석결과 남성(1.04 1.04)이 여성(0.88, 0.88)에 비해 평균적으로 좋은 시력을 보였다. 몸무게와 키의 경우 각 각 점차 증가하는 추세를 보이고 있다. 이는 남성이 여성에 비해 키가 크고 몸무게가 많이 나가기 때문에 증가하는 추세처럼 보이는 것으로 판단된다. 음주 및 흡연 여부에서도 역시 음주를 하는 집단과 흡연을 하는 집단이 그렇지 않은 집단에 비해 더 높은 평균 시력을 보이고 있는데, 이 또한 음주와 흡연 비율이 높은 남성이 많기 때문에 나타난 수치라고 판단된다. 따라서 음주 및 흡연이 시력과의 직접적인 연관성이 있다고는 보기 힘들다고 판단된다. 나이의 경우 30대 후반부터 시력이 급격히 감소하는 것을 알 수 있다. 이는 노안(원시안) 및 노화 현상으로 급격하게 시력이 저하 되고 있다고 예상된다.

Ⓒ 혈당 분석

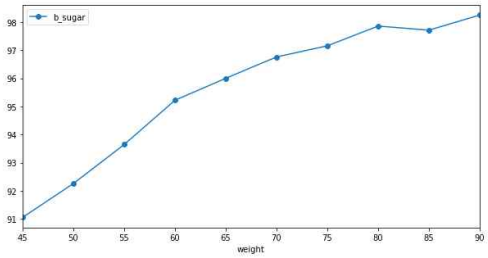
성별 vs 혈당



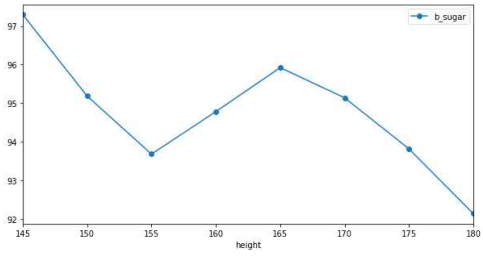
나이 vs 혈당



몸무게 vs 혈당



키 vs 혈당

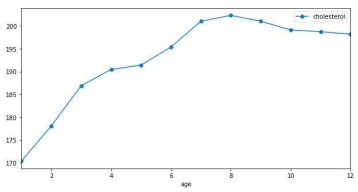


혈당의 경우 공복 혈당과 식후 2시간 혈당을 측정하여 수치를 파악한다. 본 건강 검진 데이터에서 측정한 혈당은 공복 혈당이다. **공복 혈당**의 경우 8시간 이상 금식 후 측정한 혈당 농도이다. 126mg/dL 이상이면 당뇨병, 100~125mg/dL 이면 공복 시 포도당 장애로 의심한다. 100mg/dL 이하를 정상 수치로 판단한다.

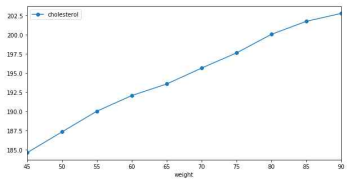
분석 결과 성별, 나이, 몸무게 그리고 키에서 혈당과의 상관관계를 보였다. 먼저 성별을 분석한 결과 남성(96.25)이 여성(93.00)에 비해 평균적으로 다소 높은 혈당 수치를 보였다. 나이를 분석한 결과 연령이 증가함에 따라 혈당이 점차 증가함을 알 수 있는데 특히 60세 이상(101.01)에서 평균적인 수치가 포도당 장애로 나타나는 만큼 당뇨병 위험 군에 속하는 사람들이 많음을 알 수 있다. 몸무게를 분석한 결과 역시 몸무게가 증가함에 따라 혈당 수치가 증가하고 있음을 알 수 있다. 그러나 그 수치가 100mg/dL를 넘지 않는 것으로 보아 몸무게와 당뇨가 양의 상관관계를 가지고 있다고 보기는 어렵다고 판단된다. 키를 분석해본 결과 감소 추세를 보였다가 다시 증가 추세를 보이고 감소를 하는데 이는 155cm ~ 160cm 사이에 여성의 비율이 높고 165cm ~ 170cm 사이에 남성의 비율이 높아 혈당의 평균적인 수치가 여성에 비해 높은 남성이 많은 구간에서 혈당이 높은 것으로 판단된다. 전반적으로는 그 이상으로 큰 키를 가졌을 때 혈당이 점점 낮아지는 것으로 예상된다.

ⓓ 콜레스테롤 분석

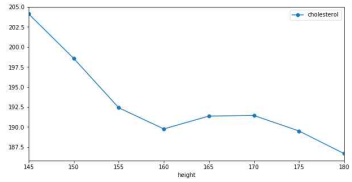
나이 vs 콜레스테롤



몸무게 vs 콜레스테롤



키 vs 콜레스테롤



콜레스테롤의 경우 총콜레스테롤, 중성지방, HDL 콜레스테롤, LDL 콜레스테롤 총 4가지 지표로 수치를 측정한다.

총콜레스테롤은 혈중 콜레스테롤의 총합으로 98 ~ 199mg/dL를 정상범위로 본다. 그 이상인 경우 동맥경화나 관상동맥질환의 위험도가 높아진다.

중성 지방의 경우 혈중 지질의 한 형태로 10 ~ 149mg/dL를 정상범위로 본다. 그 이상인 경우 마찬가지로 동맥경화나 관상동맥질환의 위험도가 높아진다.

HDL 콜레스테롤은 혈중 지질의 한 형태로서 남성의 경우 35 ~ 55mg/dL, 여성의 경우 45 ~ 65mg/dL를 정상범위로 본다. 이 콜레스테롤의 경우 좋은 콜레스테롤로서 수치가 높을수록 동맥경화나 관상동맥질환의 위험도를 낮춰준다.

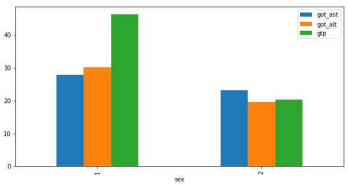
LDL 콜레스테롤은 0~130mg/dL를 정상 범위로 본다. 이 콜레스테롤은 HDL과 반대로 나쁜 콜레스테롤로서 수치가 높을수록 동맥경화나 관상동맥질환의 위험도가 높아진다.

2005년 데이터의 경우 HDL 과 LDL을 측정하지 않았기 때문에 총콜레스테롤 수치로 분석하였다. 분석 결과 나이, 몸무게 그리고 키에서 콜레스테롤과의 상관관계를 보였다. 나이의 점차 콜레스테롤 수치가 증가함을 보이는데, 특히 40세 이상부터 빠르게 증가하며 50세

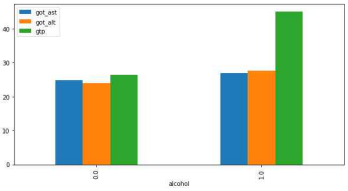
이상 65세 이하에서 평균 콜레스테롤 수치가 200mg/dL 이상으로 가장 높아 동맥경화나 관상 동맥질환의 위험성에 많이 노출되어 있음을 알 수 있다. 몸무게를 분석해 보면 몸무게의 증가에 따라 콜레스테롤 수치가 증가함을 알 수 있는데 80kg(200.01)이상에서 위험 수치에 포함되었다. 키를 분석해 보면 키가 증가함에 따라 감소하는 추세를 보이는데 이때 150cm 이하에서 200g/dL 이상의 수치를 보이는데 정확한 이유를 알 수는 없었다.

Ⓔ 간 기능 분석

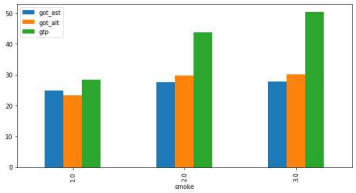
성별 vs 간 기능



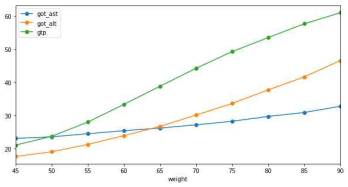
음주여부 vs 간 기능



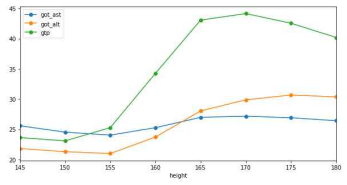
흡연여부 vs 간 기능



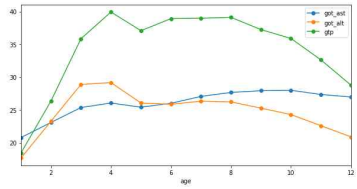
몸무게 vs 간 기능



키 vs 간 기능



나이 vs 간 기능



간 기능은 AST, ALT 그리고 γ-GTP 값에 의 총 3가지 지표로 수치를 측정한다.

AST는 간 기능을 나타내는 수치로 간세포 이외에 심장, 신장, 뇌, 근육 등에도 존재하는 효소로 이러한 세포들이 손상을 받는 경우 농도가 증가하게 된다. 0 ~ 33 IU/L를 정상범위로 본다.

ALT는 간 기능을 나타내는 수치로 간세포 내에 존재하는 효소로 이러한 세포들이 손상을 받는 경우 농도가 증가하게 된다. 0 ~ 38 IU/L를 정상범위로 본다.

γ-GTP는 간 기능을 나타내는 수치로 간 내의 쓸개관에 존재하는 효소로 글루타민산을 외부에 팹티드나 아미노산 등으로 옮기는 작용을 한다. 쓸개즙 배설 장애 및 간세포 장애가 발생할 경우 혈중에 증가하게 된다. 남성의 경우 11 ~ 56 IU/L 그리고 여성의 경우 8 ~35 IU/L를 정상범위로 본다.

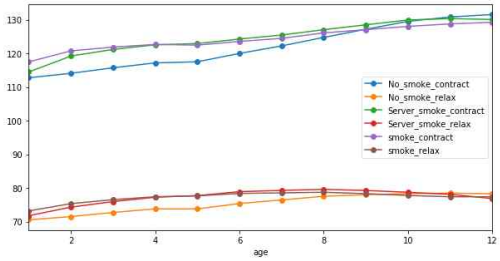
분석 결과 도시를 제외한 나머지 요소들에서 간 기능과의 상관관계를 보인다. 성별을 보면 남성이 여성에 비해 높은 수치를 보이고 있다. 이는 음주여부와 흡연여부에서도 확인 할 수 있는데 음주와 흡연을 하는 사람들이 평균적으로 그렇지 않는 사람들에 비해 간수치가 월등히 높음을 알 수 있다. 몸무게를 보면 몸무게가 증가함에 따라 간수치가 점차 증가함을 볼 수 있는데, 몸무게 85kg 이상은 평균 수치가 기준 범위를 넘어 선다는 점에서 각별히 주위를 기울여야 함으로 판단된다. 키를 보면 165cm ~ 175cm 사이에서 높은 수치를 보이는데 이는 해당 구간에 남성의 비율이 높기 때문에 수치가 높은 것으로 판단된다. 나이를 보면 30대에서 60대 구간에서 간수치가 높게 나타나는데 이는 흡연과 음주 비율이 가장 높은 나이대가 존재하는 구간으로 그것이 영향을 미친 것으로 판단된다.

◆ 2변수(기본 정보)에 대한 건강 상태 경향 분석

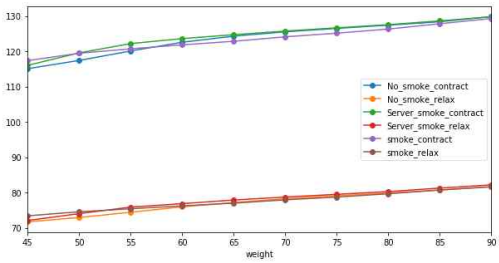
각 연도별 분석 결과 경향 추이가 모두 동일하게 나타나 2009년도 데이터를 기반으로 분석 결과를 설명하겠다. 각 항목별로 특정 2변수에 대해서만 설명하겠다.

Ⓐ **혈압 분석**

나이&흡연여부 vs 혈압



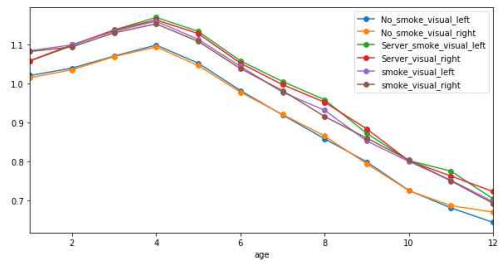
몸무게&흡연여부 vs 혈압



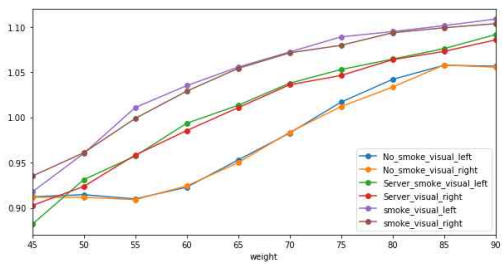
분석 결과 나이에 따라 혈압은 점차 상승하지만 동일한 나이 대에 대하여 흡연자 또는 흡연 경험자가 미흡연자에 비하여 높은 혈압 수치를 보인다. 특히 흡연자의 경우 30대 이상부터 고혈압 위험성을 가지고 있다. 흥미로운 점은 60대 이상 부터는 흡연 여부와 관계없이 혈압 수치가 비슷한 양상을 나타낸다. 몸무게를 보면 혈압은 몸무게가 증가함에 따라 점차 증가 하고 있다. 그래프에서 보면 미흡연자가 흡연자에 비해 혈압이 낮은 듯 보이나 65kg이상에서는 흡연 여부와 관계없이 동일하게 증가하는 양상을 보인다.

Ⓑ **시력 분석**

나이&흡연여부 vs 시력



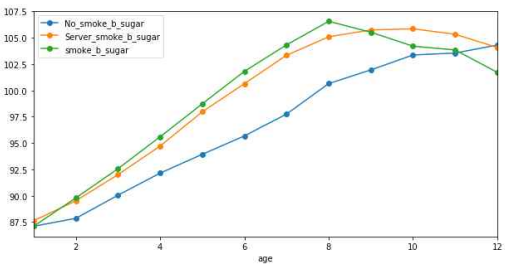
몸무게&흡연여부 vs 시력



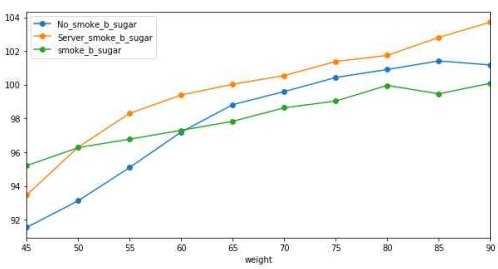
분석 결과 30대 후반부터 시력이 급격이 감소함을 보이고 있다. 흡연 여부에 관계없이 비슷한 양상을 보인다는 점에서 흡연이 시력에 영향을 미친다고 보기는 힘들다. 몸무게를 보면 전체적으로 몸무게가 증가함에 따라 시력이 증가하고 있음을 보인다. 그러나 그 이유는 그래프만으로는 설명하기 힘들다. 흥미로운 점은 흡연자가 비흡연자들에 비해 시력이 평균적으로 높다는 점이다. 앞서 단일 변수에서 남성이 여성에 비해 평균적으로 시력이 높기 때문에 상대적으로 남성의 비율이 높은 흡연자들의 시력이 높다고 판단되어지나 구체적인 이유를 설명하기에는 부족하다.

Ⓒ **혈당 분석**

나이&흡연여부 vs 혈당



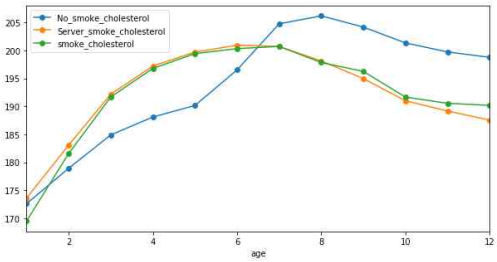
몸무게&흡연여부 vs 혈당



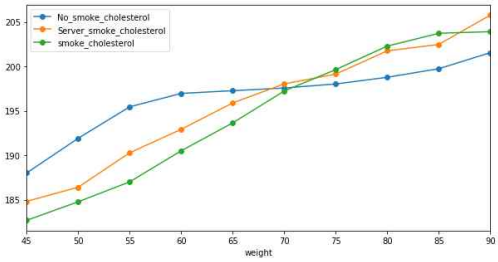
분석 결과 나이가 증가함에 따라 혈당 수치가 점차 증가함을 보이고 있다. 이 그래프에서는 흡연자가 비흡연자에 비해 훨씬 혈당 수치가 높다는 것을 알 수 있다. 흡연자의 경우 40세 이상의 경우 기준 정상 수치인 100mg/dL을 넘어서는 것을 알 수 있는데, 흡연자가 당뇨병에 노출될 가능성이 더욱 높아짐을 알 수 있다. 몸무게를 보면 몸무게가 증가함에 따라 혈당 수치는 점차 증가하는 것으로 나타난다. 흥미로운 점은 흡연자보다도 흡연 경험이 있었던 사람들이 흡연자와 비흡연자에 비해 혈당 수치가 더욱 높음을 알 수 있었다.

ⓓ 콜레스테롤 분석

나이&흡연여부 vs 콜레스테롤



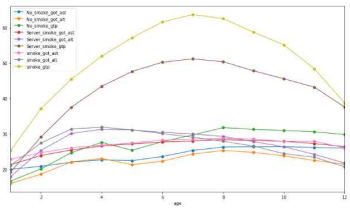
몸무게&흡연여부 vs 콜레스테롤



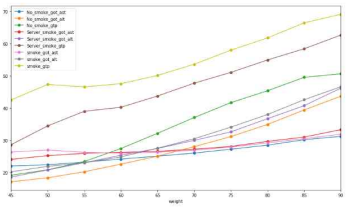
분석 결과 콜레스테롤 수치는 나이가 증가함에 따라 점차 증가하다가 감소하는 추이를 보이고 있다. 흥미로운 점은 비흡연자의 경우 흡연자에 비해 낮은 콜레스테롤 수치를 나타내었지만 40세 이상 부터는 오히려 흡연자에 비해 더욱 높은 콜레스테롤 수치를 나타내었다. 몸무게를 보면 몸무게가 증가함에 따라 콜레스테롤 수치가 증가하는 추세를 보인다. 흥미로운 점은 70kg 이상의 경우 콜레스테롤 수치가 정상 범위를 넘어서려 하는데 이때 흡연자가 비흡연자에 비해 더욱 급격하게 콜레스테롤 수치가 높아지는 것을 알 수 있다.

ⓔ 간 기능 분석

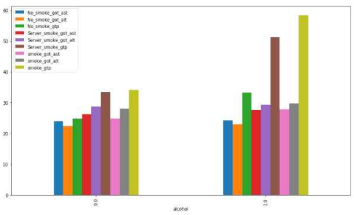
나이&흡연여부 vs 간 기능



몸무게&흡연여부 vs 간 기능



음주여부&흡연여부 vs 간 기능



분석 결과 45세에서 54세 사이에 흡연을 하는 사람들의 간수치가 기준치를 넘어서는 것을 알 수 있다. γ-GTP값을 기준으로 비흡연자는 동일 나이 대에 30IU/L 대를 보이는 반면에 흡연자의 경우 60IU/L 이상의 수치를 보이고 있다. 몸무게를 보면 몸무게가 증가함에 따라 간수치가 점차 증가함을 보이는데 이 역시 흡연자들이 비흡연자들에 비해 매우 높은 수치이다. 특히 75kg이상의 흡연자들은 동일 몸무게를 가진 비흡연자들에 비해 간수치가 기준치를 넘어섰으며 간 기능이 매우 좋지 않음을 알 수 있다. 더불어 음주여부와 흡연여부를 고려하여 분석해 보았을 때, 음주와 흡연을 모두 하는 사람들의 간수치가 가장 높은 것을 알 수 있었다. 이를 통해 음주와 흡연 모두 간 수치에 악영향을 준다는 것을 알 수 있다.

※ 본 보고서에서 분석하지 않는 나머지 2변수 경향에 대한 그래프는 jupyter notebook 파일로 대체하겠다.

◆ 데이터 학습

한 해의 수진자의 기본 정보(성, 연령, 거주지), 신체 정보(키, 몸무게), 음주여부와 흡연 여부를 n 개씩 묶어서 x 변수로 지정 한 후 그에 대한 건강 상태를 y로 두어 train파일을 만든 후 머신러닝을 을 시도해 보았고, 다른 해의 기본 정보를 test파일로 만들어 test 파일의 y 값을 예측하고 이를 실제 값과 비교해 보았다. 다음은 jupyter notebook에서 머신러닝을 실행해본 결과다.

train set

```
train=GJ_2005[['sex','age','city','height','weight','smoke','alcohol','visual_left','visual_right','b_press_contract','b
train.head()
```

	sex	age	city	height	weight	smoke	alcohol	visual_left	visual_right	b_press_contract	b_press_relax	cholesterol	b_sugar	got_ast	got_alt	gtp
0	2	1	29	145	40	1.0	0.0	1.0	1.2	100	60	165	96	12.0	21.0	13
1	2	1	41	145	40	1.0	0.0	0.9	0.9	110	75	176	99	17.0	8.0	8
2	2	1	27	145	40	1.0	0.0	1.0	1.0	116	68	139	81	13.0	9.0	13
3	2	1	47	145	40	1.0	1.0	1.2	0.8	120	70	138	72	24.0	18.0	33
4	2	1	41	145	40	1.0	1.0	0.5	0.8	120	80	119	87	11.0	14.0	14

test set

```
test=GJ_2006[['sex','age','city','height','weight','smoke','alcohol','visual_left','visual_right','b_press_contract','b_
test.head()
```

	sex	age	city	height	weight	smoke	alcohol	visual_left	visual_right	b_press_contract	b_press_relax	cholesterol	b_sugar	got_ast	got_alt	gtp
0	2	1	44	145	35	1.0	0.0	1.2	1.2	100	70	145	89	17.0	10.0	9
1	2	1	26	145	40	1.0	0.0	0.9	1.0	100	70	179	71	18.0	11.0	11
2	2	1	44	145	40	1.0	1.0	0.6	0.7	130	80	178	80	19.0	16.0	14
3	2	1	11	145	40	1.0	1.0	1.5	1.1	119	77	138	82	22.0	11.0	14
4	2	1	43	145	45	1.0	0.0	1.2	1.5	100	80	141	85	23.0	14.0	11

linear regression

Ridge

```
from sklearn.linear_model import Ridge
```

```
ridge=Ridge(alpha=0.01).fit(X_train,Y_train)
```

```
score_train=ridge.score(X_train,Y_train)
score_train
```

```
0.29521145530783532
```

```
score_test=ridge.score(X_test,Y_test)
score_test
```

```
0.28986677890664891
```

Lasso

```
from sklearn.linear_model import Lasso
```

```
lasso=Lasso(alpha=0.0001,max_iter=100000).fit(X_train,Y_train)
```

```
score_train=lasso.score(X_train,Y_train)
score_train
```

```
0.29521132794018912
```

```
score_test=lasso.score(X_test,Y_test)
score_test
```

```
0.28987399436185879
```

Decision tree

```
from sklearn.tree import DecisionTreeRegressor
```

```
tree=DecisionTreeRegressor().fit(X_train,Y_train)
```

```
score_train=tree.score(X_train,Y_train)
score_train
```

```
0.32890497834634488
```

```
score_test=tree.score(X_test,Y_test)
score_test
```

```
0.29010058451305287
```

선형 회귀 분석과 Decision tree 회귀 분석을 사용하여 train file을 만들었지만 test 파일로 예측해본 결과 score가 0.3 이상으로 증가하지 않았다. x 변수를 달리해보면서 계속 시도해 보았지만 쉽지 않았다. 따라서 이상의 러닝을 통한 분석은 진행하지 않았다.

ⅱ. 진료 내역 데이터 분석

1. 건강 검진 데이터를 활용하여 EDA 분석을 진행하였고, 의미 있는 상관관계를 도출한 부분도 많았다. 그러나 상관관계의 일부는 상식적으로 이미 알려진 사실들이었고, 새롭게 도출한 insight의 경우 데이터만으로 설명하기는 힘들었다. 더하여 의료적인 부분이라 그와 관련된 논문도 찾아보았으나 쉽지 않았다.
2. 건강 검진 데이터를 분석한 것과 마찬 가지로 수진자의 기본 정보(성별, 나이, 도시, 검진 일자)를 변수로 하여 그에 따른 진료 과목과 주상병 코드를 분석하였다.
3. 초기 6개년 데이터를 사용하려 했으나 1개년 파일의 크기가 1GB 정도의 크기를 가져 6개의 데이터셋을 활용하기에 어려움이 많았다. 따라서 2013년부터 2015년 사이의 3개년 데이터를 활용하여 분석을 진행하였다.
4. 기본 정보를 월별, 나이&성별, 도시별로 분리하여 각각 진료 과목과 주상병 코드를 분석하였다.
5. 분석은 총 2가지 방향으로 진행 되었다. 첫 번째는 진료 과목 별 환자 분포 히스토그램을 그리고 가장 환자들이 많이 찾는 진료과목 상위 5개 과목에 대해서 분석을 진행하였다.

6. 두 번째는 주 상병코드를 통해 환자들이 가장 많이 찾는 중 상병 코드 상위 20개를 뽑아 그 중 첫 번째 분석에서 진행했던 주요 진료과목에서 진료를 보는 주요 질병관련 상위 항목 3가지를 선정하여 분석하였다.

7. 각 연도별의 경향 추이를 보려는 것이 아니었기에 2013년도부터 2015년도까지의 진료내역 데이터를 합하여 3개년 데이터를 merging하여 사용하였다.

(a) 월별 분석

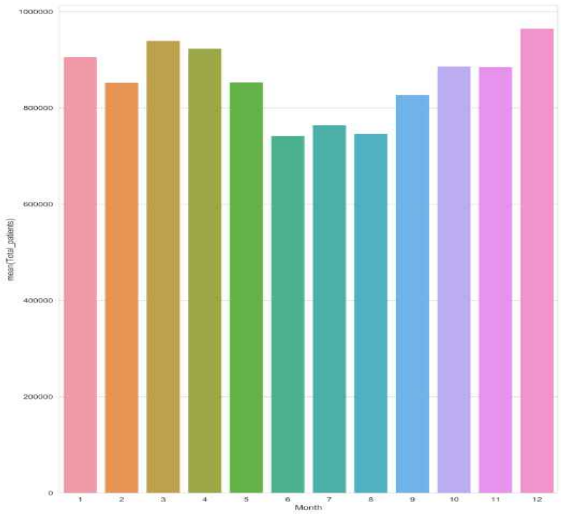
◆ 분석 방법

진료 과목 중에서 사람들이 가장 많이 찾았던 상위 5개의 진료과목에 대해 분석을 진행했다. 해당 과목에 대해서 월 별로 어느 월에 사람이 제일 많이 찾는지를 알아보았다. 또한 가장 많이 내원했던 상위 3개의 월에 대해서 구체적으로 주상병 코드를 분석했다. 주상병 코드 또한 상위 3개의 항목만 선별해서 내과에 특정 월에 찾은 환자 의 상위 3개의 특정 질병을 살펴보았다.

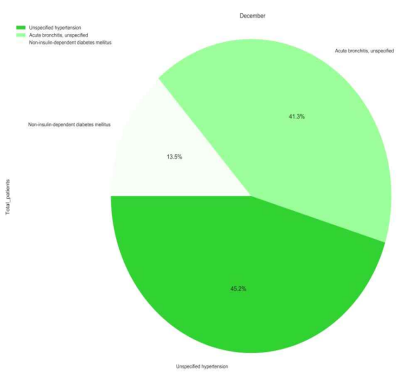
◆ 분석 내용

ⓐ 월 별 내과 이용인원 및 상위 3개에 대한 주상병 코드 분석

	Month	Total_patients
0	1	905777
1	2	852628
2	3	939279
3	4	923154
4	5	852992
5	6	741851
6	7	764242
7	8	746298
8	9	827153
9	10	886332
10	11	885001
11	12	964808

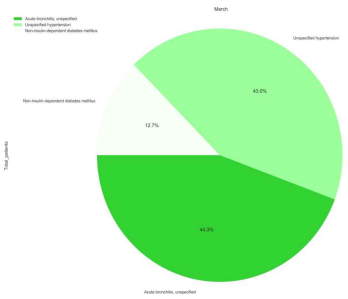


내과의 월별 인원을 불러 온 결과 위와 같았다. 오른쪽 그래프의 x 축은 월을 나타내고 있고 y 축은 Total Patients를 나타내고 있다. 전체적으로 고른 분포를 보였지만 유독 사람들은 12월, 3월, 4월에 많았고 6월~8월은 환자 수가 현저히 적었다. 생각해본 이유로 뒤에 주상병 코드를 분석하면서 좀 더 확고해졌지만 감기 환자가 겨울에 많기 때문에 이러한 그래프가 나왔다고 생각할 수 있었다. 이에 우리는 특정 월 12, 3, 4월에 대해서 주상병 코드를 살펴보았고 이를 파이 그래프로 나타내었다. 또한 각 각의 주상병 코드를 분석할 때 상위 3개의 주상병 코드를 살펴보았다.



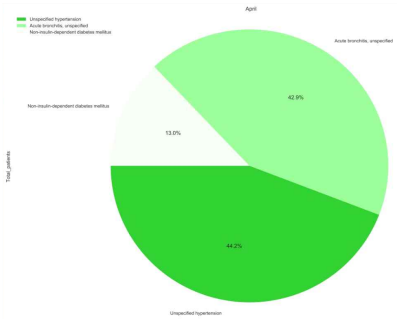
옆의 파이그래프는 12월에 해당하는 상위 3개의 주상병 코드이며 제일 큰 비율을 차지했던 질병은 고혈압이었으며 그 다음은 기관지염 마지막으로 당뇨병이었다. 상당히 많은 인원이 12월에 고혈압으로 내과를 많이 내원했으며 기관지 염 또한 많은 이유로 찾은 것을 볼 수 있었다.

12월	- I109 기타 및 상세 불명의 원발성 고혈압
	- J209 상세 불명의 급성 기관지염
	- E119 합병증을 동반하지 않은 2형 당뇨병



엮의 파이그래프는 3월에 해당하는 상위 3개의 주 상병 코드이며 제일 큰 비율을 차지했던 질병은 기관지 염이었다. 그 다음은 고혈압이었으며 마지막으로 당뇨병이 3위를 차지했다. 위와 비교해 본다면 상당히 유사한 질병으로 내과를 내원하는 환자수가 많은 것을 알 수있었다.

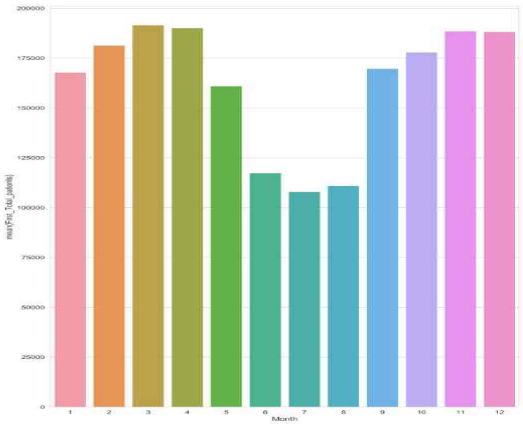
3월	- J209 상세 불명의 급성 기관지염
	- I109 기타 및 상세 불명의 원발성 고혈압
	- E119 합병증을 동반하지 않은 2형 당뇨병



엮의 파이그래프는 4월에 해당하는 상위 3개의 주상병 코드이며 제일 큰 비율을 차지했던 질병은 고혈압이었으며 그 다음은 기관지염 마지막으로 당뇨병이었다. 상당히 많은 인원이 12월에 고혈압으로 내과를 많이 내원했으며 기관지 염 또한 많은 이유로 찾은 것을 볼 수 있었다.

4월	- I109 기타 및 상세 불명의 원발성 고혈압
	- J209 상세 불명의 급성 기관지염
	- E119 합병증을 동반하지 않은 2형 당뇨병

이를 통해 내과에서는 사람이 많이 찾는 12월에 고혈압 환자를 대상으로 하는 치료 서비스나 복지 서비스를 구축한다면 좀 더 많은 사람들에게 좋은 서비스를 제공할 수 있는 것으로 보인다. 또한 기관지염으로 찾는 환자수가 많기 때문에 상위 3개 월에서 약국이나 주변 상점과 서로 힘을 맞춰서 그들을 위한 마스크나 약을 좀 더 할인된 가격이나 많이 구비해 놓으면 서로에게 좋은 마케팅 전략을 세울 수 있을 것으로 보인다. 위와 같은 방법으로 나머지 4개의 진료 과목에 대해서도 분석해 보았다. 전체적으로 12월에 환자수가 제일 많이 병원을 내원했고 특이하게 정형외과와 안과는 7~8월에 사람들이 병원을 많이 내원했다. 그 이유는 여름에 운동이나 야외 활동이 많아지기 때문에 사람들이 활동하는 활동량이 많다고 생각했다. 또한 야외 활동으로 인해 눈에 이물질, 결막염등이 많아져서 7월~8월에 환자수가 제일 많다고 결론을 지었다. 따라서 병원의 마케팅 전략에서는 7~8월에는 안과나 정형외과 같은 과의 진료시간을 늘린 다던가 좀 더 많은 수요를 맞출 수 있다면 좋은 경영전략이 될 수 있을 것 같다.
다음으로 주상병코드의 순위를 매겨 상위 5개의 항목에 해당하는 주상병코드를 대상으로 어떤 월에 많이 찾는지에 대해 살펴보았다. 이 과정에서 우리는 제일 처음 해당 주상병코드에 대해 지역별로 환자수를 나누었고 제일 환자수가 많았던 경기도 지역에 대해서 월 별 분석을 진행했다.



옆의 그래프를 보면 경기도의 감기 환자수는 3월, 4월, 12월이었고 이는 위에서 분석 했던 내과의 월 별 환자수와 상당히 유사했다. 이와 같이 식도염, 근골격계 환자수를 경기 지역에 한에서 월 별로 살펴보았다. 그런데 감기 환자의 경우 내과만 내원하는 것이 아닌 이비인후과와 소아청소년과에서도 많이 내원을 하는 것을 위의 분석을 통해 알게 되었기 때문에 분석을 하는데 있어서 좀 더 세부적으로 나눠서 분석을 진행한다면 좀 더 정확한 정보를 알 수 있을 것이다. 또 한 지역과 월 별의 변수를 통해 특정 지역에서 제공할 수 있는 데이터를 한정할 수 있을 것 같다.

(b) 성별 및 연령 분석

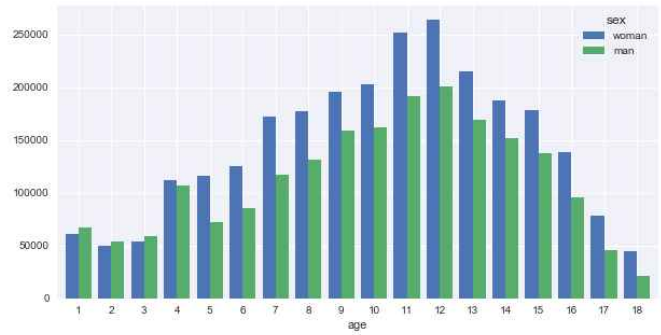
◆ 분석 방법

3개년 파일 병합 시 한 사람이 같은 주상병 코드를 가지고 같은 진료과목에 여러 번 내원한 것은 한 번 내원한 것으로 바꿔서 분석을 진행하였다.

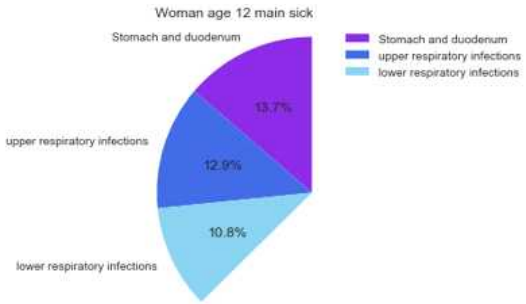
각 진료과목마다 성별과 연령대별로 얼마나 많은 사람들이 내원하는지 알아보았고 순위를 매겨서 가장 많이 내원하는 연령대 1위와 2위를 찾아보았다. 1위와 2위에 대해서는 내원하는 이유를 주상병 코드로 알아보았다. 이 때, 각 진료과목에 내원하는 성별과 연령대의 절대적인 수가 중요하기 때문에 성별, 연령대별 사람 수의 차이는 고려하지 않았다. 또한, 가장 많이 나타나는 주상병 코드가 무엇인지 알아보고 1위부터 5위까지 순위를 매겨서 각각의 주 상병 코드는 어떤 연령대와 성별이 가장 많이 걸리는지 알아보았다.

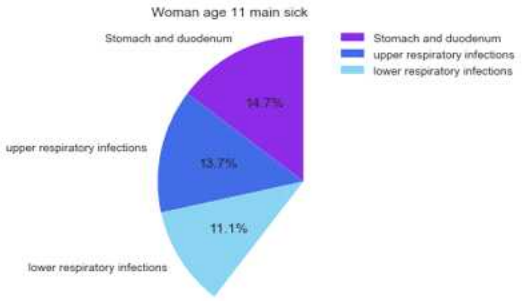
◆ 분석 내용

내과에 진료이력이 있는 사람을 대상으로 살펴본 결과는 다음과 같다. 파란색그래프는 여성, 초록색그래프는 남성이며 x축은 연령대이고 y축은 사람 수이다.

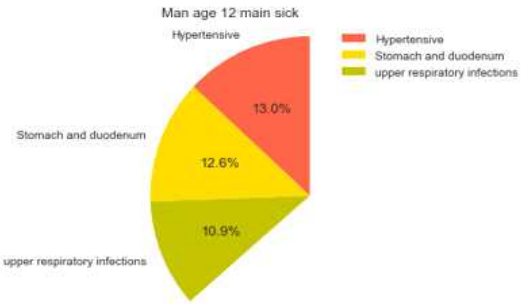


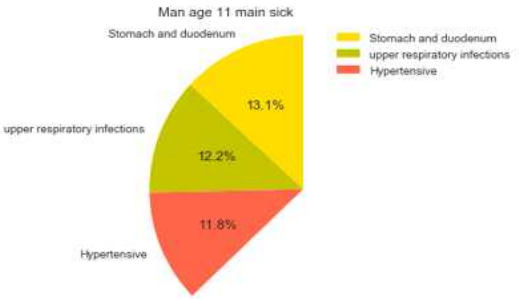
전체적으로 여성의 수가 더 많았고, 남성과 여성 모두 연령대 11(50~54세)와 12(55~59세)의 내원자 수가 가장 많았다. 연령대 11과 12의 남녀의 주상병 코드를 살펴보았고 그 중 가장 많은 3개의 주상병 코드의 비율을 계산하여 파이 그래프로 나타내었다.





왼쪽이 연령대 12 여성의 주상병 1위부터 3위이고 오른쪽은 연령대 11 여성의 주상병을 나타낸 것이다. 1위는 위, 식도 및 십이지장 질환이었고 2위와 3위는 상기도 및 하기도감염의 감기였다.





남성의 경우도 왼쪽이 연령대 12, 오른쪽이 연령대 11의 주상병코드를 1위부터 3위까지 나타낸 것이다. 붉은 색이 고혈압, 노란색이 위,식도 및 십이지장 질환이고 초록색이 상기도 감염, 감기를 나타낸다. 이를 통해 50세에서 60세 사이의 사람들이 내과를 가장 많이 방문하는 것을 알 수 있었다. 내과를 방문하는 대부분의 이유는 감기와 위, 식도 질환이라는 것을 알 수 있었고 여성의 경우 위,식도 및 십이지장 질환이 많았다. 남성의 경우에는 고혈압이 3위 안에 들어있는데, 50세에서 55세 사이에서는 고혈압이 3위지만 56세 이상이 되면 고혈압이 1위로 올라가는 것을 알 수 있었다.

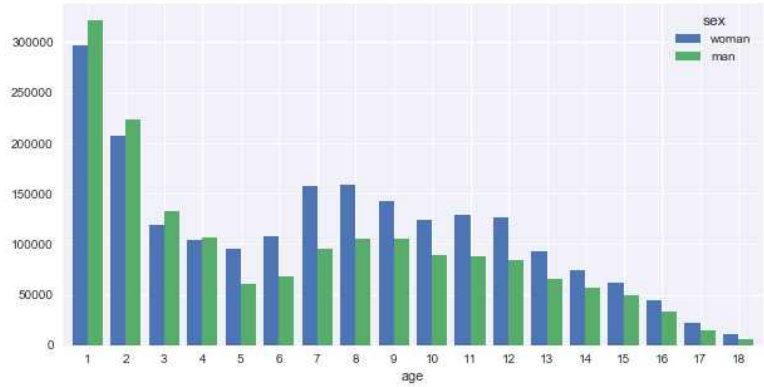
따라서 내과에서는 50세와 60세를 대상으로 고혈압과 위, 식도 및 십이지장 질환을 예방하고 치료하는데 도움을 주는 서비스를 제공하거나 감기 환자의 잦은 방문에 대비하여 필요한 물품을 구비해 놓거나 감기 환자를 위해 마스크 등의 도움이 되는 물품을 서비스를 제공하면 좋을 것이라는 마켓팅 전략 및 병원 경영 전략을 세울 수 있을 것이다.

위와 같이 다른 4개의 진료과목에 대해서도 분석해보았다. 전제적으로 대부분의 진료과목에 대해서 여성의 수가 남성의 수보다 많았고 보통 연령대 11과 12가 가장 많았다. 주상병코드에 대해서는 상위 1위와 2위의 연령대가 대부분 같은 주상병코드를 가지고 있었고 남성과 여성간의 주상병 코드도 비슷했다.

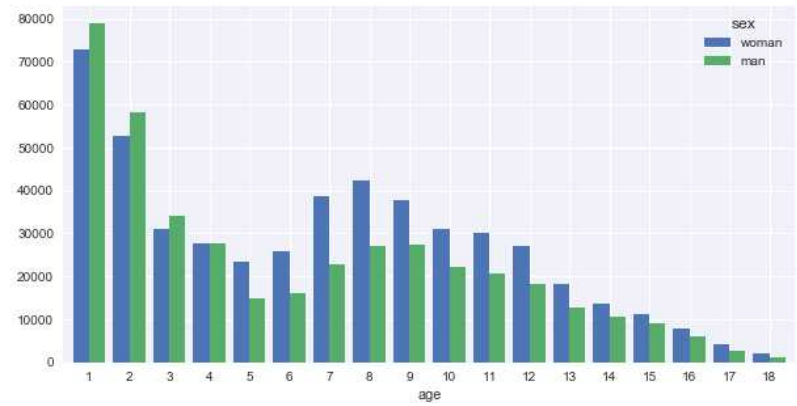
특별한 경우를 몇 가지 이야기하면 먼저, 소아청소년과의 경우는 전체적으로 남성의 내원수가 더 많았고 연령대 4(15~19세)부터는 내원수가 1만명 이하로 줄어든다. 연령대 1(0~4세)의 수가 가장 많았고 두 번째로 많았던 연령대 2(5~9세)보다 그 수가 2배정도 많았다. 주상병코드는 상기도 및 하기도 감염과 질환으로 모두 감기와 관련된 코드였다. 이비인후과의 경우, 여성은 연령대 1(0~4세), 8(35~39세)가 가장 많았고 남성은 연령대 1(0~4세), 2(5~9세)가 가장 많았으며 주상병코드는 귀의 질환보다도 감기와 관련된 질환의 비율이 더 많았다.

이를 통해 병원이 마켓팅을 할 때, 어떤 연령대와 성별을 대상으로 하면 더 효과적일지 알 수

있을 것이고, 소아청소년과는 영유아들의 감기 예방접종에 관한 알림 서비스를 하면 좋을 것이다 등의 마켓팅 아이디어를 내는데 도움이 될 것이다.
다음은 주상병코드의 수로 순위를 매겨보았을 때, 1부터 5위에 해당하는 주상병코드를 대상으로 어떤 성별과 연령대가 많이 걸리는지 알아보았다. 예를 들어 감기에 대한 분석그래프는 다음과 같다.



이처럼 위 및 식도 질환, 관절과 연조직 장애 그리고 피부염에 대해서 성별, 연령대별 내원수를 분석해보았다. 이를 통해 주상병코드별로 어떤 연령대와 성별이 많은지 알 수 있었고 전국적으로 이 4가지의 주상병코드를 가지고 내원하는 사람의 수가 굉장히 많기 때문에 이 사람들을 수요자로 하는 어떤 물품이나 서비스를 제공할 수 있을 것이다. 또한, 이들을 대상으로 하는 캠페인이나 서비스를 계획한다면 그 대상의 연령대나 성별에 따라 마켓팅 전략을 세울 수 있을 것이다. 위의 결과가 지역별로 어떻게 나타나는지도 궁금해졌고 감기, 위 및 식도 질환 그리고 관절과 연조직 장애 각각의 내원수가 가장 많은 지역을 찾았고 그 지역에서의 성별, 연령대별 내원 수를 분석해보았다. 세 가지 질병 모두 경기도 지역의 내원수가 가장 많았다. 예를 들어 경기도 지역의 감기환자에 대해 분석한 그래프는 다음과 같다.



경기도 지역의 그래프 추이는 위에 나타냈던 전국적인 감기환자의 그래프와 비슷한 추이를 보였다. 이를 통해 위와 비슷하게 경기도의 감기환자에게 어떤 서비스를 제공한다면 그 대상을 어떤 연령대와 성별로 해야 효과적인지 알 수 있을 것이다. 감기환자가 내원하는 병원은 내과, 소아청소년과와 이비인후과가 있었는데 추가적으로 세 병원을 변수로 추가해서 분석한다면 병원별로 어떤 연령대와 성별의 감기환자들이 많이 내원하는지를 알 수 있을 것이다. 추가적으로 경기도의 연령대별, 성별 분포 데이터가 있다면 이를 통해 경기도에서 병원이 많이 필요한 지역이 어딘지, 감기환자가 많을 것으로 예상되는 지역은 어딘지, 지역의 감기환자 연령대 및 성별에 따라 내과, 소아청소년과와 이비인후과의 비율은 어떻게 달라져야하는지 그리고 그 비율에 따라 어느 지역이 내과가 부족하거나 많은지를 알 수 있을 것이다.

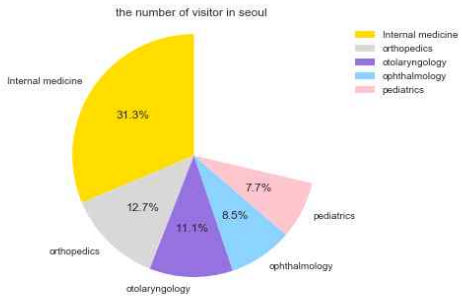
(c) 시·도 별 분석

◆ 분석 방법

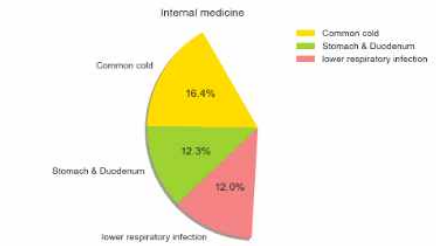
각 도시별로 주요 5개 진료과목 방문 비율을 분석하였다. 이후 각 도시별 상위 진료과목 4개에 대하여 어떠한 이유로 진료를 받기 위해 병원을 방문하는지를 알아보기 위하여 주 상병 코드를 분석하고 주요 원인 3개에 대하여 분석하였다. 다음으로 주요 질병 군 3개에 대하여 각 도시별 분포를 그리고 해당 도시의 인구와 해당 도시의 주요 질병을 진료하는 진료과목 수를 활용하여 수용 가능 환자수를 모델링하고 현재 병원수가 부족한지 풍부한지를 실제 환자수와 예상 환자수를 비교하여 분석하였다.

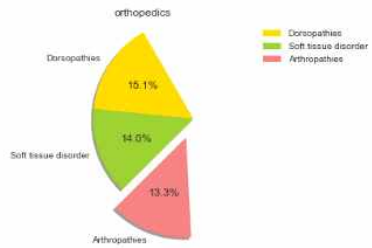
◆ 분석 내용

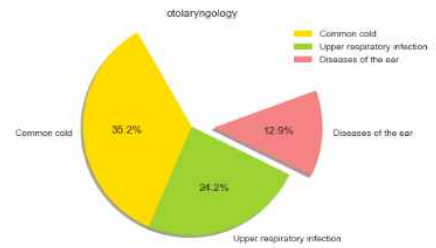
세종 특별시를 제외한 나머지 특별시, 광역시 그리고 각 도별로 분석을 진행하였다. 서울을 예로 들면 다음과 같다.

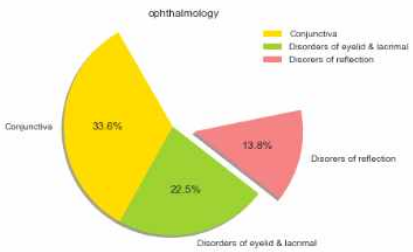


서울시의 경우 많이 찾는 병원은 내과, 정형외과, 이비인후과, 안과 그리고 소아 청소년과 이다. 전국 데이터와 비교하면 서울시에서는 소아 청소년과 방문 비율이 상대적으로 작 은 것으로 보인다. 이 그래프를 바탕으로 상위 4개의 진료과목에서 어떠한 질병으로 병원을 방문하는지를 분석하였다.



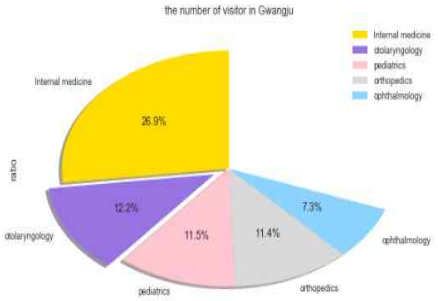




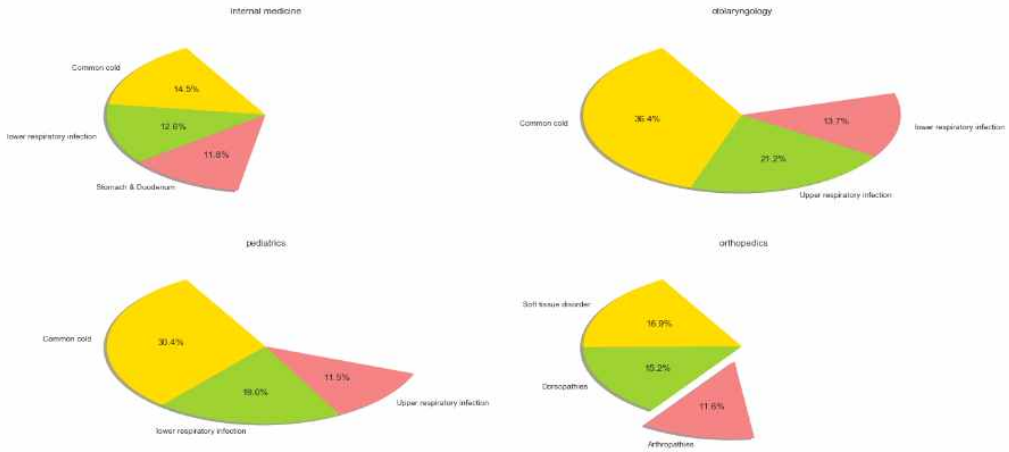


내과의 경우 일반 감기, 복통 및 위염 그리고 하기도 감염(기관지 염) 등의 이유로 병원을 방문하였다. 정형외과의 경우 주로 가벼운 타박상, 연조직 장애 그리고 관절염으로 병원을 방문하는 것으로 나타났다. 이비인후과 역시 감기, 상기도 감염(기관지 염) 그리고 중이염 등의 이유로 방문하는 것으로 나타났다. 안과의 경우 결막염, 눈꺼풀 및 눈물기관 장애 그리고 반사 장애등의 이유로 병원을 방문하는 것으로 나타났다.

일반적인 도시와는 다른 특징을 보이는 도시를 예로 들어 보겠다.

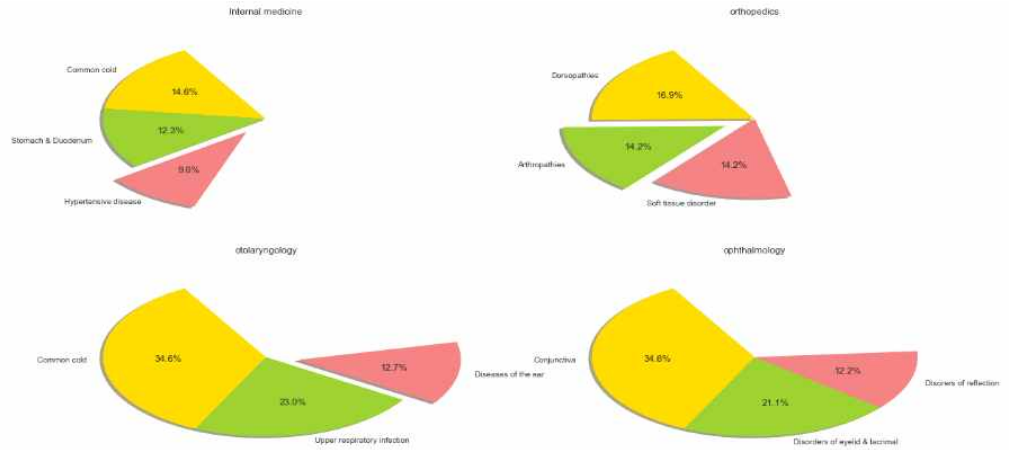


다른 도시들은 내과 다음으로 정형외과를 많이 찾는 반면에 광주는 많은 사람들이 이비인후과를 찾는 것으로 나타났다. 그 다음으로는 소아 청소년과의 비율이 많았다. 각 항목별로 구체적으로 병원을 찾는 주요 원인을 분석해 보면 아래와 같다.



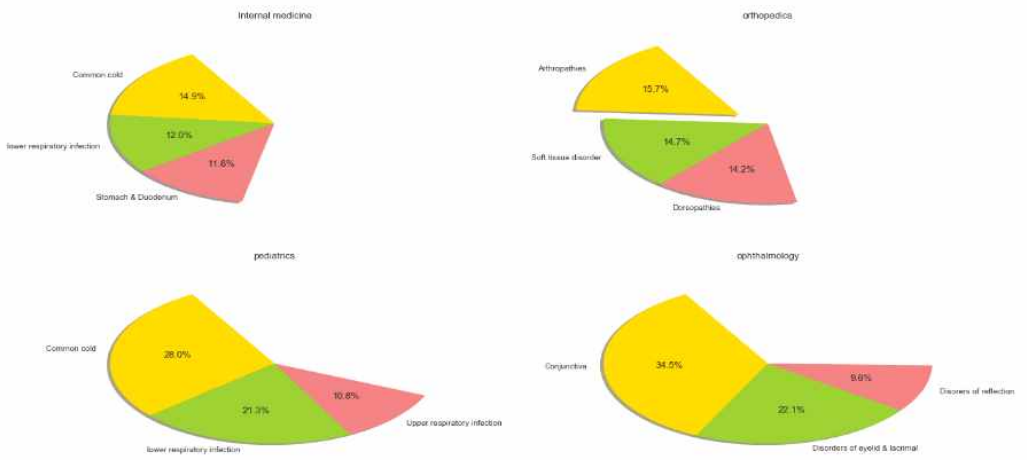
주요 원인은 다른 도시와 크게 다르지 않았다. 내과 이비인후과는 감기와 기관지염으로 많이 찾았고, 소아 청소년과 역시 대부분 감기와 기관지염으로 병원을 찾았다. 정형외과 역시 연조직 장애, 타박상 그리고 관절염으로 많이 찾았다.

강원도의 경우는 다른 도시와 다르게 흥미로운 점을 발견하였다.

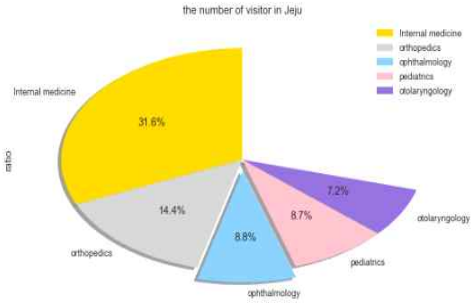


병원을 찾는 이유가 여타 도시들과 크게 다를 바가 없어 보이나 내과를 방문하는 사람들 중 많은 사람들이 고혈압성 질병으로 찾는 것을 확인할 수 있다. 이는 이전 건강 검진 데이터에서도 확인할 수 있는데 각 시도별 평균 혈압이 강원도에서 상위권에 속하는 것을 알 수 있다. 의료계에서는 이점에 집중하여 강원도에서 고혈압성 환자들이 상대적으로 인구대비 많은 이유를 찾고 그것을 타겟으로 마케팅 전략을 세운다면 좋을 것이라 예상한다.

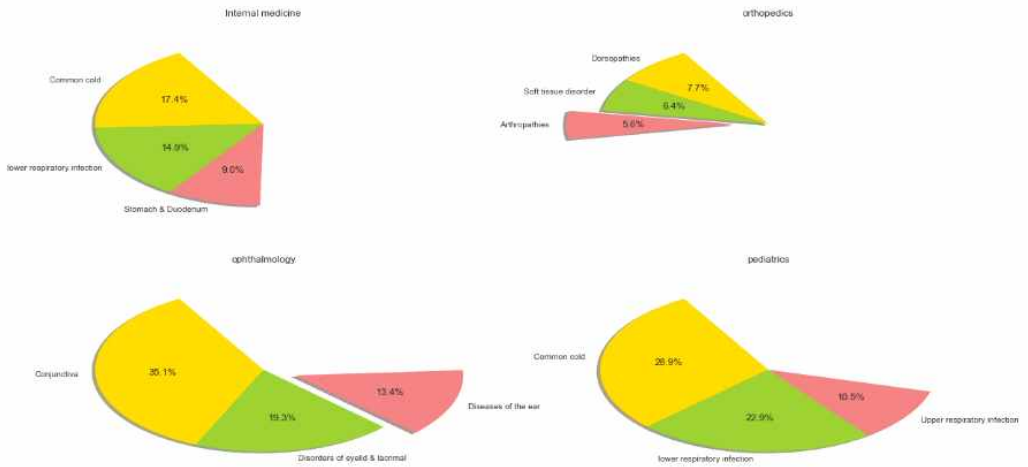
경상북도에서도 흥미로운 결과를 확인 할 수 있었다.



다른 도시와는 다르게 경북 지역에서는 정형외과를 방문하는 이유 중에 관절염이 가장 크게 차지하였다. 그 원인을 분석해 보면, 타 지역에 비해서 경북지역에 노령인구 층이 상대적으로 높아서 대다수 노령층의 주요 질병인 관절염이 많은 것이 아닐까 예상 한다. 의료 마케팅 전략을 세울 때 경북 지역에서 정형외과를 연다면 노인층을 대상으로 관절염 전문 병원을 많이 증설하거나 관절염 환자를 위한 여러 물리 치료 시설 및 설비 등을 구축하는 것을 고려해 볼 필요가 있다고 판단된다.



제주도 지역에서는 다른 지역들과 다르게 안과를 많이 찾는 것을 알 수 있다. 구체적인 관련성을 찾기는 힘들지만 북부 지역에서 남부 지역으로 내려올 수 록 안과를 찾는 비율이 점차 증가하는 추세를 감안해 볼 때 이 부분도 고려해보아야 할 부분이라고 생각한다.

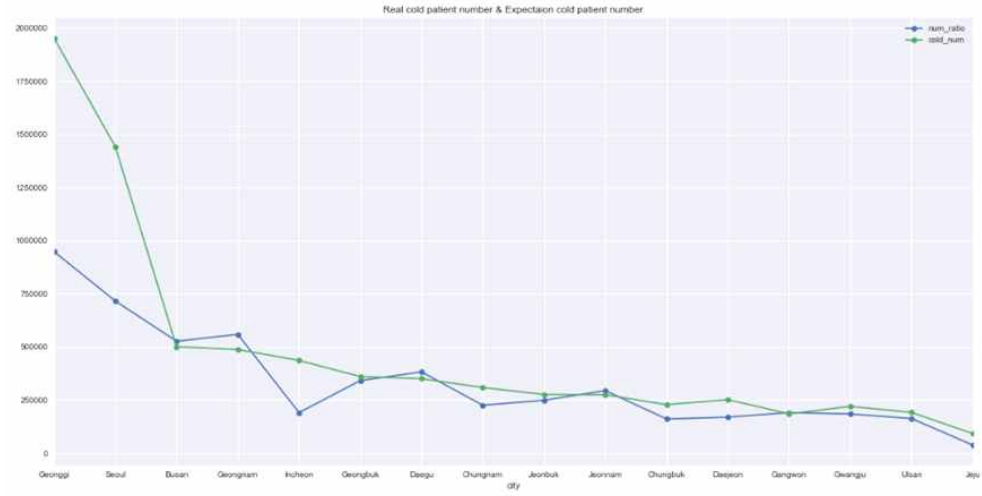


구체적으로 살펴보면 더욱 흥미로우면서도 이상한 점을 발견 할 수 있었다. 제주도에서 안과를 찾는 주요 원인 중에 중이염이 있다는 사실이다. 그것도 전체의 13%나 차지 한다는 점에서 굉장히 흥미로운 결과였다. 그 원인을 정확히 알 수는 없지만 만약 데이터를 만들 당시 오류가 전혀 없었다고 가정한다면 그 이유를 면밀하게 조사해 보는 것도 좋을 것 같다.

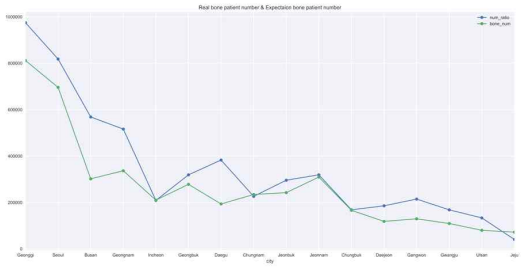
주 상병 코드를 중심으로 주요 질병에 대하여 각 도시별 환자수를 분석해 보았다.
먼저 감기, 상·하기도 감염 환자를 대상으로 도시별 실제 환자수를 나타내었다. 그리고 이 환자들이 질병을 치료하기 위해 방문하는 진료 과목을 찾은 결과 95%가 내과, 이비인후과, 소아청소년과를 방문하였다. 이를 통해 각 도시별 내과, 이비인후과, 소아과의 수를 고려하여 해당 질병을 치료하는 목적으로 각 진료과를 방문하는 비율을 계산하여 예상되는 도시별 환자 수를 모델링 하였다. 다음은 모델링한 식이다.

$$Expectation\ patient\ number = 3 * 2880 * \left(N_i * R_i + N_p * R_p + N_o * R_o\right)$$

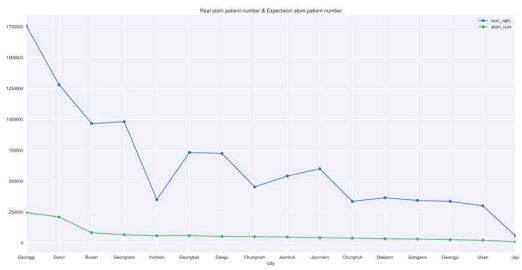
이때, 3은 3개년 데이터를 사용하였기에 3년을 의미하며, 2880은 의사 한명이 1년간 진료를 맡을 수 있는 환자 수를 산출 한 것이다. 산출 근거로는 의사가 주 5일 근무하며 하루 평균 6시간 동안 진료를 보며 시간 당 2명의 환자를 본다고 가정 하였다. 이후 각 진료과목에 해당 진료과목의 방문 비율을 곱하여 예상 환자수를 만들었다. 아래 그래프는 예상 환자 수(푸른색 그래프)와 실제 환자 수(초록색 그래프)를 나타낸 것이다. 가령 경기도의 경우 예상된 환자 수에 비하여 실제 환자수가 매우 많은 것으로 보아 내과, 이비인후과 그리고 소아 청소년과의 추가 증설이 필요한 것으로 예측하였다.



관절·척추 및 연조직 질환



식도, 위 및 십이지장 질환



관절·척추 및 연조직 질환 그래프의 경우 실제 환자 수에 비해 예상 환자수가 동일하거나 적기 때문에 현재 정형외과로도 충분히 커버가 가능하다고 판단된다. 식도, 위 및 십이지장 질환은 대다수 내과를 통해 진료를 받는데 실제 환자수에 비해 예상 환자수가 매우 큰 것으로 보아 현재로도 충분하다는 것을 알 수 있다.
물론 이와 같은 예측에 여러 가지 문제점들이 있다. 첫째, 모델링에서 제시한 의사 한명당 커버 가능한 환자 수를 산출한 근거에서 값이 하나라도 변화하게 되면 위 그래프와 전혀 다른 결과가 나올 수 있다. 실제 위 모델이 사용되기 위해서는 그 점에 대하여 더욱 깊이 있게 고민

해보아야 할 것이다. 둘째, 감기 환자의 경우 3개의 진료과목을 묶어 산출 하였기 때문에 실제 어떠한 진료 과목을 증설해야 하는지를 바로 알아 볼 수가 없다. 따라서 각 과목별로 분리하여 동일 질병에 대하여 필요한 병원 수를 다시 모델링 할 필요가 있다. 셋째, 가령 경기도 감기 환자의 경우 실제 환자수가 예상 환자 수에 비해 훨씬 많다는 점에서 병원이 많이 부족한 것으로 판단 할 수 도 있겠지만 가령 그 지역의 의료 체계가 잘 되어 있어 적은 의사로도 많은 환자들을 커버 할 수 있는 가능성 또한 배제 할 수 가 없다. 이 외에도 위 모델을 사용하기 위해 고려해야 할 상황들이 많을 것으로 보인다. 이후 모델을 위한 데이터 들을 추가 수집하여 만들어 낸다면 더욱 정확한 모델이 되지 않을까 생각한다.

Ⅳ. 토의

처음 목표는 건강검진데이터를 이용하여 수진자의 기본정보, 신체정보, 음주 및 흡연 여부와 각 건강검진 항목별 값들 사이의 유의미한 관계를 찾아내는 것이며, 이를 통해 기본정보가 주어지면 그와 유사한 집단을 통해 건강검진 결과를 예측하는 것이었다. EDA방법을 통해 분석하던 과정에서 유의미한 관계를 찾아도 그에 대한 원인이나 근거를 주어진 데이터만 가지고는 찾을 수 없다는 문제점을 발견했다. 단순히 현재 일어나고 있는 현상을 보여줄 순 있었지만 이 현상이 다음에도 똑같이 반복될지, 어떤 요소에 크게 영향을 받는지, 받지 않는지 등을 알 수 없었고 따라서 이 결과를 응용해서 또 다른 유의미한 결과를 도출하기 어려웠다. 또한 특정 인물의 건강검진결과와 유사집단의 건강검진결과는 차이가 많아서 예측이 의미가 없다고 판단하였다. 오히려 사람들의 질병에 주목해보면 어떨까라는 의견이 나왔고 진료내역데이터를 이용하는 것으로 방향을 바꾸게 되었다.

◆ 건강검진 데이터를 분석하며...

> 건강 검진 데이터 분석에서 생각보다 다양하고 엄청난 부류로 나누어서 EDA를 진행했는데 생각보다 흥미롭고 유의미한 정보를 끌어낼 수 없었다. 그 이유는 단순성이 제일 크다고 생각했다. 누구나 다 생각해 낼 수 있는 데이터 분석은 좋은 분석이라고 생각할 수 없었기에 건강 검진 데이터 분석이 좋다고 생각할 수는 없었다. 하지만 긍정적인 부분은 연도별로 EDA를 진행하면서 경향성을 볼 수 있었고 필요하고 필요 없는 데이터를 골라내고 분석하는 데에 있어서 의의를 두었다.

> 40대를 지나 50대로 갈수록 좋은 것은 줄어들고 나쁜 것은 잘 없어지지 않는다는 것을 알게 되었다. 예를 들어 시력은 계속 나빠지고 콜레스테롤은 그대로 라는 것. 데이터를 정제할수록 경향성이 보이는 것이 신기하면서 한편으로는 혹시 정제하는 과정해서 무엇이 누락되거나 어떤 요소가 고려되지 않아서 이런 결과가 나온 것은 아닌지 의심만 늘어갔다. 다음에는 확률적인 검증의 방법을 익혀서 의심을 해소시키고 싶다.

> 머신러닝을 통해 기본 정보를 활용하여 건강 상태를 예측해보자는 큰 포부로 시작하였지만 어려운 일이 많았다. 건강이라는 요소는 단순히 나의 키와 성별과 몸무게 등만으로 결정 되는 것이 아니기 때문이다. 나의 식습관 그리고 운동 여부 등 다양한 개인적인 요소에 의해 만들어지는 것이기 때문이다. 그러나 각 기본 요소별로 특징들을 분석하는 일은 상당히 흥미로웠다. 다만 의료적인 지식이 부족하여 멋진 insight를 도출하기 힘들었다는 점이 가장 아쉽다.

◆ 진료내역 데이터를 분석하며...

> 3000만개의 데이터를 합칠 때 컴퓨터가 터질 뻔 했던 것처럼 내 머리도 터질 뻔 했다. 데이터의 시각화가 굉장히 중요하다는 것을 깨달았는데 늘어져있는 코딩 글자들을 정신없이 쳐다보다가 갑자기 알록달록한 파이 그래프가 그려지면 그 순간은 나도 모르게 눈이 초롱초롱해졌다. 물론 데이터의 경향성을 파악할 때에도 많은 도움을 주었다.

> 진료 내역 데이터 분석은 생각보다 흥미로운 점이 많았다. 건강 검진 데이터와는 다르게 실질적인 병명과 진료과목 코드를 확인할 수 있었기 때문이다. 환자들은 과연 어느 달, 어느 연령대, 어느 지역에서 특정 진료과에 많이 찾고 특정 질병을 가졌는지 확인하고 분석하는 작업은 건강 검진보다 많은 의미를 끌어낼 수 있었다고 생각한다. 또한, 점점 분석을 좁혀가면서 특수성까지 높였기에 좀 더 의미 있는 데이터 분석이 아니었나 생각된다.

> 약 1GB 짜리의 데이터를 직접 이용해 본 것은 처음이다. 사실 오직 글자로 된 문서가 그렇게 커다란 용량으로 존재 할 수 있다는 사실 조차 몰랐다. 시간이 많았다면 더욱 다양한 분석을 할 수 있지 않을까 아쉬움이 남는다. 이후 기회가 된다면 동일 데이터를 활용하여 더욱 정확하게 모델링을 하고 또한 의료 산업에 정말 도움을 줄 수 있는 의료 마케팅 전략서를 만들고 싶다.

◆ 분석의 활용 방안 및 아쉬운 점
분석에서 이야기 했듯이, 내과는 50세에서 60세까지 가장 많이 찾고 그 이유는 위 및 식도 질화, 고혈압과 감기 때문이다. 따라서 내과에서는 이들을 대상으로 고혈압과 위, 식도 및 십이지장 질환을 예방하고 치료하는데 도움을 주는 서비스를 제공하거나 감기 환자의 잦은 방문에 대비하여 필요한 물품을 구비해 놓거나 감기 환자를 위해 마스크 등의 도움이 되는 물품을 서비스를 제공하는 등의 마켓팅 전략 및 병원 경영 전량을 세울 수 있을 것이다.
또한 전국적으로 감기, 위 및 식도 질환과 관절질환이 병원을 찾는 이유 top3이다. 감기환자만 120만명이 넘으며 꾸준히 존재하므로 이들을 대상으로 하는 서비스나 제품을 만든다면 수요자가 부족할 일은 없을 것이다. 이 때, 어떤 성별과 연령대의 환자가 많은지 알 수 있다면 그에 따라 제품이 달라질 수도 있고 마켓팅 전략을 효과적으로 바꿀 수 있을 것이다.
마지막으로, 지역별로 어떤 연령대와 성별이 감기, 관절질환, 위 및 식도 질환에 걸리는지 알 수 있기 때문에 예를 들어, 경기도의 정형외과에서는 관절질환 때문에 어떤 연령대와 성별이 병원을 많이 찾는지 알 수 있다. 이를 통해 아이들이 많이 온다면 놀이방을 만든다던가, 노인들이 많이 온다면 등받이가 있는 의자를 준비하는 등의 해당하는 연령대와 성별을 고려해서 서비스를 제공하거나 이벤트성의 소정의 선물을 준비할 때 어떤 것이 좋을 지 등의 마켓팅 전략과 경영 전략을 세울 수 있을 것이다.
이번 데이터 분석에서는 진행하지 못했지만 건강검진 결과 데이터와 건강상태의 지표에 관한 데이터를 이용해서 특정인물의 건강검진 결과가 주어졌을 때, 유사집단의 건강상태보다 어떤 부분이 더 좋은지, 더 나쁜지, 치료나 예방은 어떻게 해야 하는지를 이야기해줄 수 있는 진단 알고리즘을 만들면 좋을 것 같다.

◆ 분석을 마치며...
> 데이터의 수가 많아서 뭐든 해보면 의미 있는 결과가 나올 것이라고 생각했는데, 건강 검진 결과를 분석한 후에 이 분석결과가 보여주는 것은 우리가 추구하는 게 아니니까 다른 방향을 설정하고 분석을 다시 해야 한다는 사실을 받아들이는 것이 어려웠다. 한편으로는 데이터 분석도 가설을 세우고 검증하고 실패하면 다시 가설을 세우고 검증하는 것이 실험과 같다는 점이 흥미로웠다.

> 가장 어려웠던 점을 꼽으면 일단 데이터의 크기가 매우 컸기에 바로바로 결과를 확인하기 매우 힘들었다. 또한, 주 상병 코드의 종류가 너무 많다 보니 분류하고 분석하기에 너무 많은 작업량을 필요로 했기에 이 부분도 어려운 부분이었다. 하지만 분석을 하다 보니 여러 가지 시각화를 통해 평소 생각지 못했던 부분에서 통상적인 생각과는 다르게 나온 데이터를 보면서 신기했고 흥미로웠다.

> 생각보다 insight를 찾는 일은 쉽지 않았다. 다양한 형태로 eda를 진행하다 보면 의미 있는 결과를 도출 시킬 수 있지 않을까 라는 막연한 기대감으로 많은 시간을 쏟아 부었지만 어려웠다. 물론 그 과정에서 배운 점도 많다. 실제 우리가 알고 있던 사실들이 데이터에서 고스란히 도출됨을 보았고, 맞추지 않은 퍼즐 조각과 같은 다양한 insight 들이 숨어 있음을 느꼈다. 앞으로 데이터 분석을 더욱 깊이 있게 공부한다면 더 많은 조각들을 맞출 수 있지 않을까 생각한다.