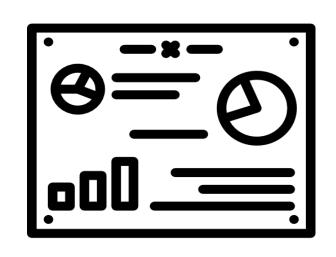
Big Data Project Final Presentation

11조

강유민, 김근하, 김대원

Topic

진료 내역 데이터를 이용한 **질병 유형** 및 **발병 경향 분석**





Goal

2013년 ~ 2015년(3년) 진료 내역 데이터를 활용하여 상위 5개 항목의 질병의 유형과 발병 경향을 분석 하고 이를 통 해 의료 산업의 마케팅 전략에 도움이 될 만한 정보를 제공

구분	의과_보건기	구분	의과_보건기
十世	관	구군	관
0	일반의	23	가정의학과
1	내과	24	응급의학과
2	신경과	25	산업의학과
3	정신과	26	예방의학과
4	외과	50	구강악안면외 과
5	정형외과	51	치과보철과
6	신경외과	52	치과교정과
7	흉부외과	53	소아치과
8	성형외과	54	치주과
9	마취통증의학 과	55	치과보존과
10	산부인과	56	구강내과
11	소아청소년과	57	구강악안면방 사선과
12	안과	58	 구강병리과
13	이비인후과	59	예방치과
14	피부과	80	한방내과
15	비뇨기과	81	한방부인과
16	영상의학과	82	한방소아과
17	방사선 종양학과	83	한방안과, 이비인
18	병리과	84	한방신경정신 과
19	진단검사의학 과	85	침구과
20	결핵과	86	한방재활의학 과
21	재활의학과	87	사상체질과
22	핵의학과	88	한방응급
ZZ	결측	-	정상 또는 해당사항 없음

Data1. 진료 내역 정보 데이터

- 국민 건강 보험 공단에서 제공한 데이터
- **연도별 데이터** 제공 (2013년 ~ 2015년)
- 국민건강보험가입자 중 해당 년도에 요양(병/의원)기관으로 부터의 진료내역이 1건 이상 있는 가입자 **100만 명을 무작위**로 선별하여 표본 형성
- 진료 내역 데이터 셋의 크기를 축소하기 위해 구간 분포 비율을 유지한 채로 정제 과정 거침

Data1. 진료 내역 정보 데이터

19 columns

	D	<u> </u>	n	-	-		ш			V		M	N	0	п	0	D	C .
조녀드	가이자 일리	진료내역일성발	# -	- 연령대코드	시드코드	9 양개시일		지르 과목 =	조산병 =	크드부상병코드	O 야 인 스							데이터 기존일
2005	800942	1	1	11		20051230	3		E14	K769	1	1	15	42770	12830	29940		20151220
2005	816566	2	2	5		20051217	3		S801	J060	1	1	15	16460	4930	11530		20151220
2005	816566	3	2	5		20051217	3		S335	S801	7	7	15	89480	21000	68480	_	20151220
2005	991692	4	2	8		20051214	3		L239	B352	1	1	15	10740	3000	7740	_	20151220
2005	385646	5	1	11		20051214	3		110	5552	1	1	15	9050	3000	6050	_	20151220
2005	795595	6	2	2	- 1	20051210	3		L239	J209	1	1	15	10740	3000	7740		20151220
2005	88924	7	2	15		20051203	3		J450	J060	3	3	15	30630	4500	26130		20151220
2005	806394	8	2	12		20051221	3		M545	J069	2	2	15	21310	6000	15310		20151220
2005	800941	9	2	10		20051206	3		M170	K31	1	1	15	11460	3000	8460		20151220
2005	789654	10	2	15		20051215	3		L239	J060	1	1	15	12380	1500	10880		20151220
2005	504830	11	2	11		20051217	3		T242		2	2		43390	13010	30380		20151220
2005	995208	12	2	3	- 1	20051231	3		J209	D212	1	1	15	59910	17970	41940		20151220
2005	23710	13	1	10	11	20051205	3	1	163	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	1508	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 10 million rows

•••

Data2. 시•도별 인구수 데이터

- 통계청에서 제공한 데이터
- **연도별 데이터** 제공 (2013년 ~ 2015년)
- **시•도별 인구수** 제공
- 데이터 활용을 위해 필요한 형태로 새롭게 가공

Data2. 시•도별 인구수 데이터

A	Α	В	С	D	Е	F	G	Н	I	
1	통계표명:	지역별 인구 및 인	구밀도							
2	단위:	천명, 명/㎢								
3			012)13	20		201		
4		민구	민구밀도	민구	민구밀도	민구	민구밀도	민구	민구밀도	
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	민천	2,794	2,684	2,830	2,718	2,862	2,722	2 883	2 748	
10	광주	1,504	3,000	1,504	3,000	1,505	3,002	A	B 005	
11	대전	1,540	2,852	1,545	2,860	1,553	2,879	1 city	number 860	
12	물산	1,125	1,061	1,137	1,073	1,151	1,085	2	29906000 097	
13	세종	102	220	118	255	132	285	3 26	10260000	
14	경기	11,974	1,177	12,126	1,192	12,282	1,207		221	
15	강원	1,504	90	1,506	89	1,510	90	4 27	7420000 90	
16	충북	1,553	210	1,565	211	1,578	213	5 28	8575000 215 256	
17	충남	2,043	249	2,062	251	2,088	254	6 29	4515000 207	
18	전북	1,817	225	1,821	226	1,829	227			
19	전남	1,782	145	1,784	145	1,792	146	7 30		
20	경북	2,656	140	2,661	140	2,671	140	8 31	3452000 141 316	
21	경남	3,265	310 303	3,278	311	3,307	314 315	9 41	36831000 324	
22 _	제주 수도권	561 24.805	2.099	570 24,946	308 2,111	583 25,119	2,124			
24			 '추계 시도편 : 2015-/		<u> </u>	25,115	2,124	10 42	4533000 <u>134</u>	
25		* 수도권 : 서울, 인		2040],五年四古	ㅜ ·시크르세1			11 43	4732000	
26			고전, 공기 구추계 시도편 : 201 5	5-2045 ° O YI⊂ A	키이그와 구든교토브		I도변 구든면전은 3	12 44	6253000	
27			구후계 시포된 : 2016 ¹ 추계는 2017년에 작		: 근 [시 시포파증구	- 시크리계리 다 사	시그는 축포단하는 /			
21		시고를 하네근도	구세도 2011 라이 그	OUNED				13 45	5485000	
								14 46	5373000	
								15 47	8010000	

16

1752000

- 2013년 ~ 2015년 3년 시도별 인구 합산!

Data3. 지역별 진료과목 데이터

- 대한 병원 협회에서 제공한 데이터
- **지역별 진료과목 현황** 제공
- 데이터 활용을 위해 필요한 형태로 새롭게 가공

- 시도, 내과, 소아 청소년과,

- 이비인후과, 정형외과, 안과

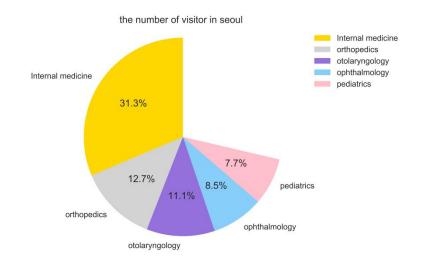
1	city	Internal m	pediatrics	otolaryngo	orthopedia	ophthalmo	ology
2	11	175	104	54	141	46	
3	26	132	77	36	98	16	
4	27	99	58	21	66	12	
5	28	48	29	12	36	6	
6	29	46	30	10	29	7	
7	30	50	17	12	32	8	
8	31	41	23	12	23	4	
9	41	240	134	67	168	3	
10	42	47	23	19	37	11	
11	43	46	21	8	29	7	
12	44	62	27	16	39	8	
13	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	

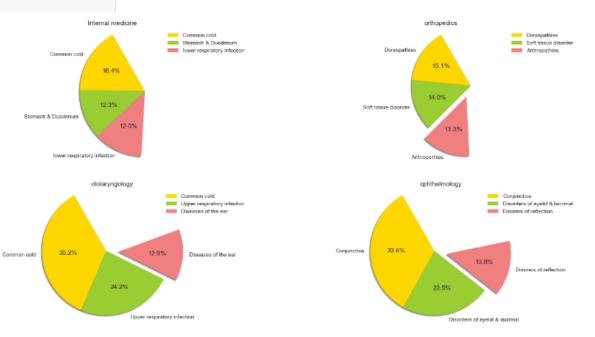
Analysis: EDA – Step 1



지역 별 분석

```
In [94]: fig = plt.figure(figsize=(20,10))
         # 4111
         data su sick 1=data su[data su['treat kind']==1]
         data su sick 1=data su sick 1.groupby('main sick s').count()
         data su sick 1['ratio']=data su sick 1['city']/data_su_sick_1['city'].sum()
         data su sick 1=data su sick 1.sort values('ratio', ascending=False)
         data su sick 1=data su sick 1.rename(index={'JO':'Common cold','K2':'Stomach & Duodenum','J2':'lower respiratory infecti
         data su sick 1=data su sick 1.iloc[:3]
         explode = (0, 0, 0)
         colors = ['gold', 'yellowgreen', 'lightcoral']
         ax1 = plt.subplot2grid((2,2),(0,0))
         plt.pie(data su sick 1.ratio, explode=explode, shadow=True, colors=colors, autopct='%1.1f%%', startangle=120, labels=data
         plt.title('Internal medicine')
         plt.axis('equal')
         plt.legend()
         # 전형외과
         data su sick 2=data su[data su['treat kind']==5]
         data su sick 2=data su sick 2.groupby('main sick s').count()
         data su sick 2['ratio']=data su sick 2['city']/data su sick 2['city'].sum()
         data su sick 2=data su sick 2.sort values('ratio', ascending=False)
         data su sick 2=data su sick 2.rename(index={'M5':'Dorsopathies','M7':'Soft tissue diso
         data su sick 2=data su sick 2.iloc[:3]
         explode = (0, 0, 0.2)
         ax1 = plt.subplot2grid((2,2),(0,1))
         plt.pie(data su sick 2.ratio, explode=explode, shadow=True, colors=colors, autopct='%1
         plt.title('orthopedics')
         plt.axis('equal')
         plt.legend()
         # OIHI인후교
         data su sick 3=data su[data su['treat kind']==13]
         data su sick 3=data su sick 3.groupby('main sick s').count()
         data su sick 3['ratio']=data su sick 3['city']/data su sick 3['city'].sum()
         data su sick 3=data su sick 3.sort values('ratio', ascending=False)
         data su sick 3=data su sick 3.rename(index={'JO':'Common cold','J3':'Upper respiratory
         data su sick 3=data su sick 3.iloc[:3]
         explode = (0, 0, 0.2)
         ax1 = plt.subplot2grid((2,2),(1,0))
         plt.pie(data su sick 3.ratio, explode=explode, shadow=True, colors=colors, autopct='%1
         plt.title('otolaryngology')
         plt.axis('equal')
         plt.legend()
        # 217
```

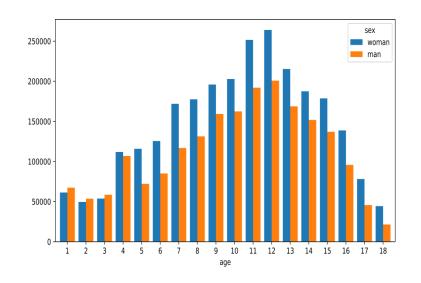




연령 & 성별 분석

```
In [33]: tw=JR_2010_2015[JR_2010_2015['treat_kind']==12]

In [32]: t_m=twe[twe['sex']==2]
    t_m=twe[twe['sex']==1]
    t_m=t_m[('sex']==1)
    t_m=t_m[('sex', 'treat_kind')]
    t_m=t_m[('sex', 'treat_kind')]
    t_m=t_m[('sex', 'treat_kind')]
    t_m=t_m(('sex', 'treat_kind'))
    t_m=t_m(('sex', 'treat_kind))
    t_m=t_m(('sex', 'treat_kind))
    t_m=t_m((sex), ('sex), ('sex))
    t_m=t_m((sex), ('sex), ('sex))
    t_m=t_m((sex), ('sex), ('sex), ('sex))
    t_m=t_m((sex), ('sex), ('sex
```

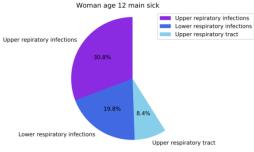


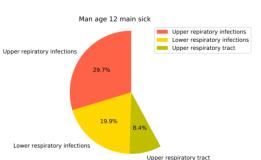
main sick by woman age 12,11

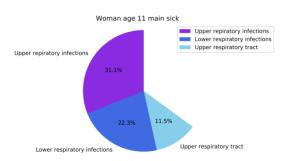


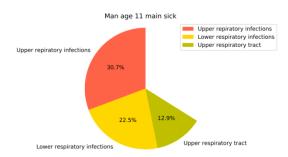
main sick by man age 2,12

```
In []: # man age 12 main sick 37#
```









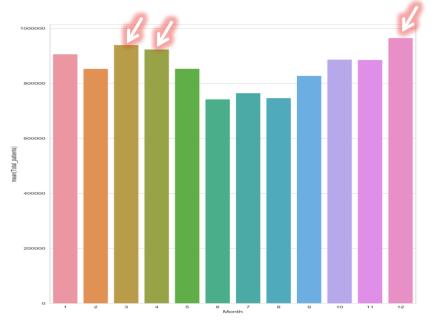
월별 분석

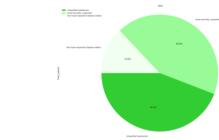
First Month Find 3 Large main sick

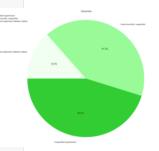
f1 = f1.rename(columns = {0: 'counts'})

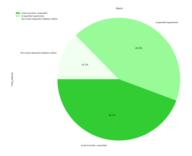
```
In [12]: al = pd.value_counts(ms_2010_2015_codel_st[ms_2010_2015_codel_st['main_sick'].str.match('A')]['main_sick'].values, sort = True)
         b1 = pd.value.counts(ms 2010 2015.code1.st[ms 2010 2015.code1.st[main.sick'],str.match('B')]['main.sick'],values, sort = True)
         c1 = pd.value_counts(ms_2010_2015_code1_st[ms_2010_2015_code1_st["main_sick"],str.match("C")]["main_sick"],values, sort = True)
         dl = pd.value_counts(ms_2010_2015_code1_st[ms_2010_2015_code1_st["main_sick"].str.match("D")]["main_sick"].values, sort = True)
          e1 = pd.value_counts(ms_2010_2015_code1_st[ms_2010_2015_code1_st['main_sick'].str.match('E')]['main_sick'].values, sort = True)
         fl = pd.value counts(ms 2010 2015 codel st[ms 2010 2015 codel st['main sick'].str.match('F')]['main sick'].values, sort = True)
          g1 = pd.value_counts(ms_2010_2015_code1_st[ms_2010_2015_code1_st["main_sick"].str.match('G')]['main_sick'].values, sort = True)
         h1 = pd.value_counts(ms_2010_2015_code1_st[ms_2010_2015_code1_st['main_sick'].str.match('H')]['main_sick'].values, sort = True)
         il = pd.value_counts(ms_2010_2015_code1_st[ms_2010_2015_code1_st['main_sick'].str.match('I')]['main_sick'].values, sort = True)
         j1 = pd.value_counts(ms_2010_2015_code1_st[ms_2010_2015_code1_st['main_sick'].str.match('J')]['main_sick'].values, sort = True)
         k1 = pd.value_counts(ms_2010_2015_code1_st[ms_2010_2015_code1_st['main_sick'].str.match('K')]['main_sick'].values, sort = True)
         11 = pd.value_counts(ms_2010_2015_codel_st[ms_2010_2015_codel_st['main_sick'].str.match('L')]['main_sick'].values, sort = True)
         m1 = pd.value_counts(ms_2010_2015_code1_st[ms_2010_2015_code1_st['main_sick'], str.match('M')]['main_sick'], values, sort = True)
         n1 = pd.value_counts(ms_2010_2015_code1_st[ms_2010_2015_code1_st["main_sick"].str.match("N")]["main_sick"].values, sort = True)
         o1 = pd.value_counts(ms_2010_2015_code1_st[ms_2010_2015_code1_st['main_sick'].str.match('0')]['main_sick'].values, sort = True)
         p1 = pd.value_counts(ms_2010_2015_codel_st[ms_2010_2015_codel_st['main_sick'].str.match('P')]['main_sick'].values, sort = True)
          q1 = pd.value_counts(ms_2010_2015_code1_st[ms_2010_2015_code1_st["main_sick"].str.match('Q')]['main_sick'].values, sort = True)
         r1 = pd.value_counts(ms_2010_2015_code1_st[ms_2010_2015_code1_st["main_sick"],str.match("R")]["main_sick"],values, sort = True)
         s1 = pd.value_counts(ms_2010_2015_code1_st[ms_2010_2015_code1_st['main_sick'].str.match('S')]['main_sick'].values, sort = True)
         t1 = pd.value_counts(ms_2010_2015_code1_st[ms_2010_2015_code1_st['main_sick'].str.match('T'))['main_sick'].values, sort = True)
         ul = pd.value_counts(ms_2010_2015_code1_st[ms_2010_2015_code1_st['main_sick'].str.match('U')]['main_sick'].values, sort = True)
         v1 = pd.value_counts(ms_2010_2015_code1_st[ms_2010_2015_code1_st['main_sick'].str.match('V')]['main_sick'].values, sort = True)
         w1 = pd.value_counts(ms_2010_2015_codel_st[ms_2010_2015_codel_st['main_sick'].str.match('\frac{\psi}{\psi})]['main_sick'].values, sort = True)
         x1 = pd.value_counts(ms_2010_2015_code1_st[ms_2010_2015_code1_st['main_sick'].str.match('X')]['main_sick'].values, sort = True)
         y1 = pd.value_counts(ms_2010_2015_code1_st[ms_2010_2015_code1_st['main_sick'].str.match('Y')]['main_sick'].values, sort = True)
          z1 = pd.value counts(ms 2010 2015 code1 st[ms 2010 2015 code1 st[main sick'],str.match('Z')]['main sick'],values, sort = True)
```

```
In [13]: code1_list = []
         a1 = pd.DataFrame(a1)
         a1 = a1.rename(columns = {D: 'counts'})
         al_max = al.iloc[al['counts'].values.argmax()]
         code1_list.append(a1_max)
         b1 = pd.DataFrame(b1)
         b1 = b1.rename(columns = {0: 'counts'})
         code1 list.append(b1.iloc[b1['counts'].values.argmax()])
         c1 = pd.DataFrame(c1)
         c1 = c1.rename(columns = {0: 'counts'})
         code1_list.append(c1.iloc[c1['counts'].values.argmax()])
         d1 = pd.DataFrame(d1)
         d1 = d1.rename(columns = {0: 'counts'})
         code1_list.append(d1.iloc[d1['counts'].values.argmax()])
         e1 = pd.DataFrame(e1)
          e1 = e1.rename(columns = {0: 'counts'})
         code1_list.append(e1.iloc[e1['counts'].values.argmax()])
         f1 = pd.DataFrame(f1)
```

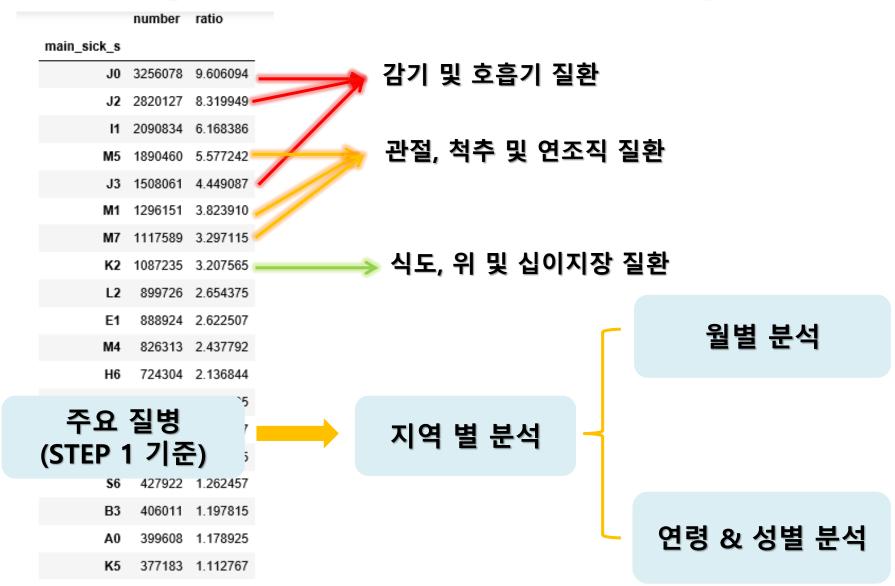


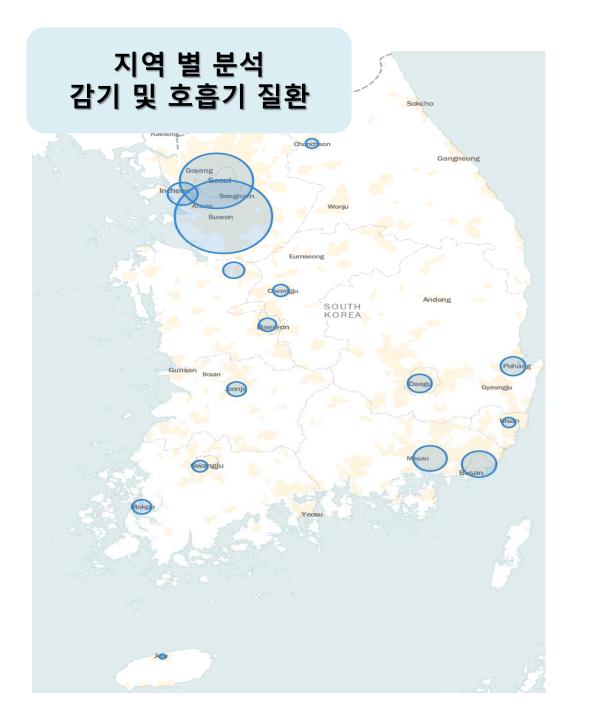






Analysis: EDA – Step 2





number

city

Geonggi	1951700
Seoul	1441390
Busan	501095
Geongnam	487949
Incheon	436450
Geongbuk	360681
Daegu	350615
Chungnam	308911
Jeonbuk	276578
Jeonnam	275215
Daejeon	252140
Chungbuk	228575
Gwangju	220330
Ulsan	192124
Gangwon	185537
Jeju	92684

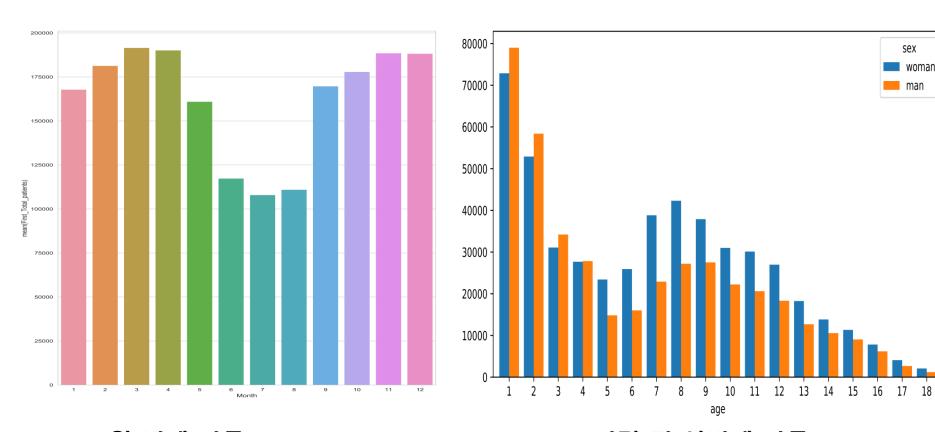
지역 별 분석 감기 및 호흡기 질환

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

Patient number per doctor in 1 year (2 patient) * (6 hour) * (5 days) * (48 weeks)



월 별 & 연령 및 성별 분석 감기 및 호흡기 질환



월 별에 따른 감기 및 호흡기 환자 수 [경기 지역]

연령 및 성별에 따른 감기 및 호흡기 환자 수 [경기 지역]

지역 별 분석 관절, 척추 및 연조직 질환 Gangneung Andong SOUTH

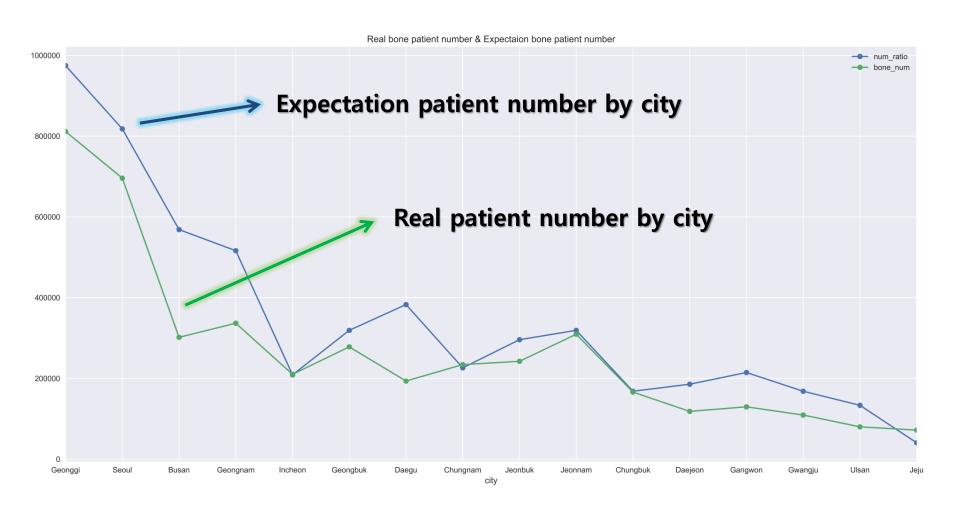
number

city

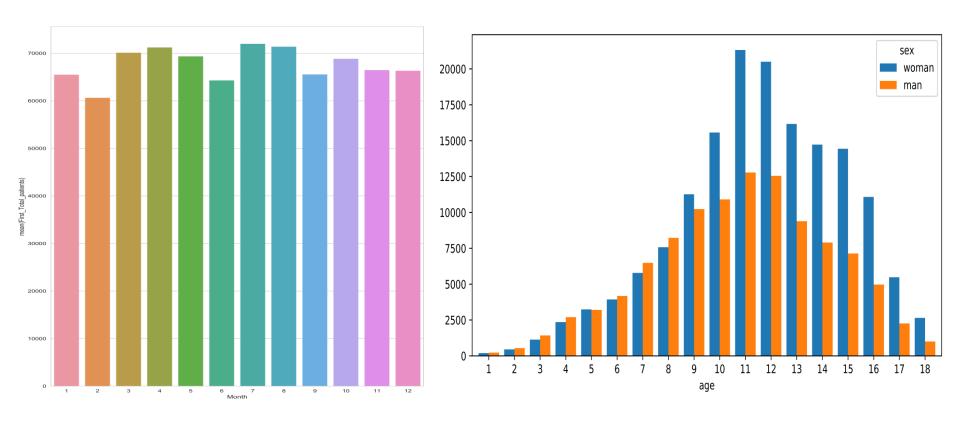
Geonggi	811525
Seoul	695991
Geongnam	336663
Jeonnam	309454
Busan	301820
Geongbuk	278134
Jeonbuk	242348
Chungnam	234223
Incheon	210143
Daegu	193588
Chungbuk	166157
Gangwon	129633
Daejeon	118360
Gwangju	109343
Ulsan	80067
Jeju	71977

지역 별 분석 관절, 척추 및 연조직 질환

Expectation patient number = $3 * 2880 * (N_o * R_o)$



월 별 & 연령 및 성별 분석 관절, 척추 및 연조직 질환



월 별에 따른 관절, 척추 및 연조직 환자 수 [경기 지역]

연령 및 성별에 따른 관절, 척추 및 연조직 환자 수 [경기 지역]

지역 별 분석 식도, 위 및 십이지장 질환 Chulcheon Gangneung Seoul Suwon Andong SOUTH Gunsan

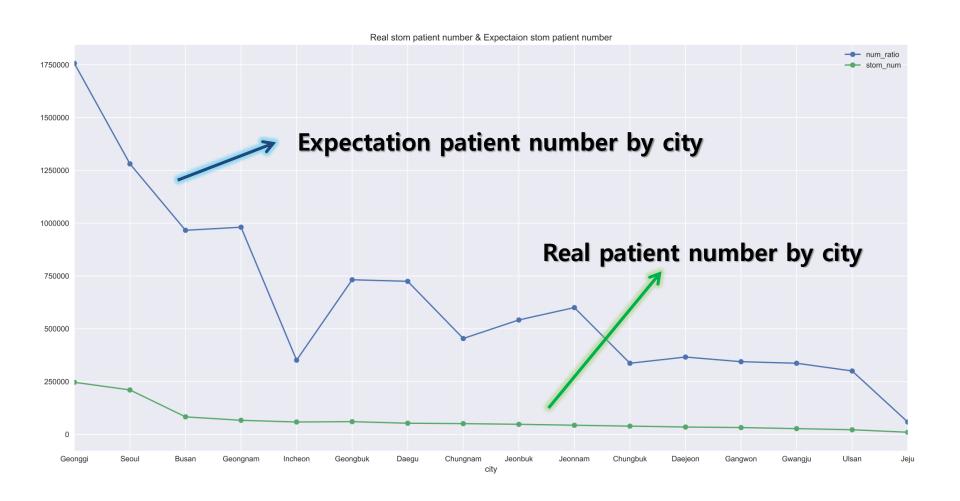
number

	100.00	1
100	ш	70.00
	ш	100
48	III S	

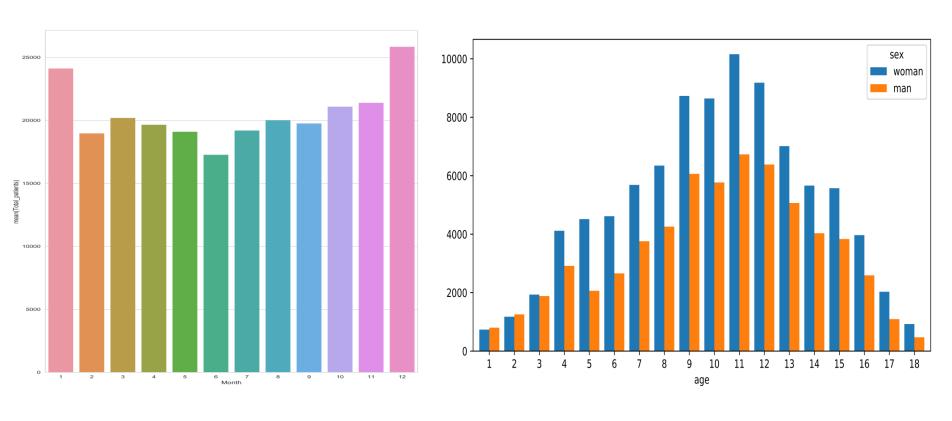
Geonggi	246536
Seoul	210314
Busan	82897
Geongnam	66491
Geongbuk	60179
Incheon	58449
Daegu	52605
Chungnam	50693
Jeonbuk	47649
Jeonnam	43136
Chungbuk	38955
Daejeon	34632
Gangwon	32161
Gwangju	27148
Ulsan	21724
Jeju	9992

지역 별 분석 식도, 위 및 십이지장 질환

Expectation patient number = $3 * 2880 * (N_i * R_i)$



월 별 & 연령 및 성별 분석 식도, 위 및 십이지장 질환



월 별에 따른 식도, 위 및 십이지장 환자 수 [경기 지역]

연령 및 성별에 따른 식도, 위 및 십이지장 환자 수 [경기 지역]

Result

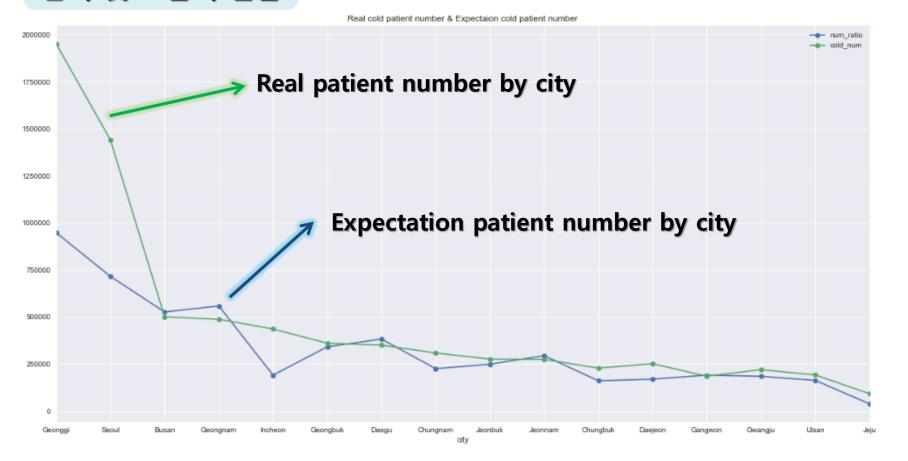
지역 별 분석 감기 및 호흡기 질환

Real patient number > Expectation patient number

⇒ Need to increase hospital

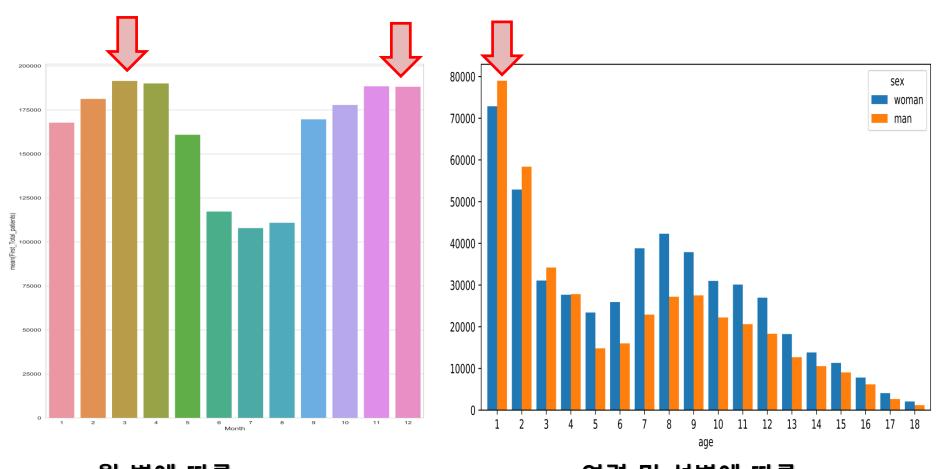
Real patient number < Expectation patient number

⇒ Exist more than enough hospital



Result

3월과 12월에 병원을 찾는 0~9세의 영유아들을 위해 병원이 그에 맞는 서비스를 제공하면 Good



월 별에 따른 감기 및 호흡기 환자 수 [경기 지역]

연령 및 성별에 따른 감기 및 호흡기 환자 수 [경기 지역]

Q&A