[비즈니스 모델링]

한국복지 패널 데이터 분석 결과 보고서

60201976 장채은

1. 데이터 소개

빈곤층, 근로 빈곤층, 차상위층의 가구 형태, 소득 수준, 취업 상태 등 이들의 계측 규모에 따른 생활 실태 변화를 동태적으로 파악한 데이터이다. 현재 본 보고서에서 사용할 16차 조사 대상 가구규모는 전체 6,240 가구이다.

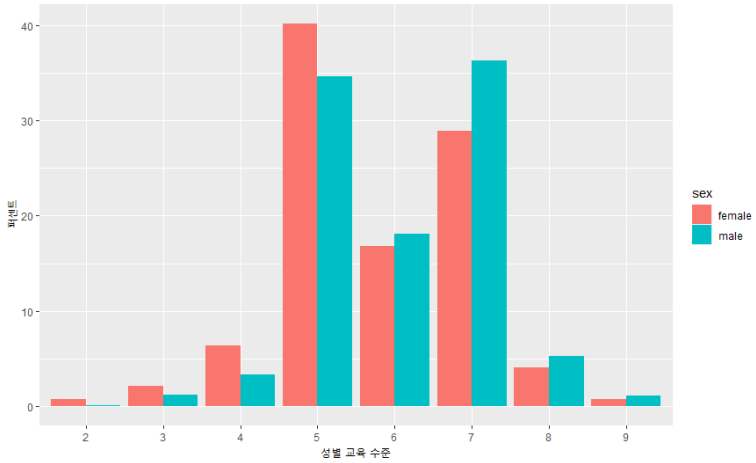
2. 데이터 분석

데이터 분석은 forecast, dplyr, ggplot2, readxl 패키지를 사용하였다. spss 데이터 파일을 불러오는 것은 forecast, 그래프를 그리는 것은 ggplot2, excel 파일을 불러오는 것은 readxl, 새로운 데이터를 만들거나 기존의 데이터를 계산하는 함수를 사용하는 것은 dplyr 패키지를 사용하였다.

(1) 성별에 따른 남녀 교육 수준 (가구원 1)

가구원1 성별에 따른 남녀의 교육 수준에 대해 분석하고자 한다. 현재 가구원에 있는 나이는 모두 성인이므로, 현재 노년층을 제외한 60대 이전의 청년, 중장년 층의 성별에 따른 남녀 교육 수준을 알아볼 것이다. 필자는 세대주를 의미하는 가구원에서 여성과 남성의 교육 수준은 크게 차이 나지 않을 것으로 예상하였다. 60대 이전의 남성과 여성의 명수는 1751: 388로 데이터의 차이가 크게 차이가 나므로 명수로 분석하지 않고 각각 성별의 명수 퍼센트로 분석한다.

	sex	graduated	n	totalSex	pct
	<chr>	<fct>	<int>	<int>	<dbl>
1	female	2	3	388	0.8
2	female	3	8	388	2.1
3	female	4	25	388	6.4
4	female	5	156	388	40.2
5	female	6	65	388	16.8
6	female	7	112	388	28.9
7	female	8	16	388	4.1
8	female	9	3	388	0.8
9	male	2	2	1751	0.1
10	male	3	21	1751	1.2
11	male	4	58	1751	3.3
12	male	5	606	1751	34.6
13	male	6	317	1751	18.1
14	male	7	635	1751	36.3
15	male	8	92	1751	5.3
16	male	9	20	1751	1.1



필요한 변수는 교육수준, 태어난 년도, 성별이며, 태어난 년도와 데이터 수집 년도(2021년)을 계산하여 나이, 연령층 변수를 새로 도출하였다. 이 변수들의 결측치와 이상치가 없기 때문에 is.na 함수를 사용하지 않고 분석을 하였다.

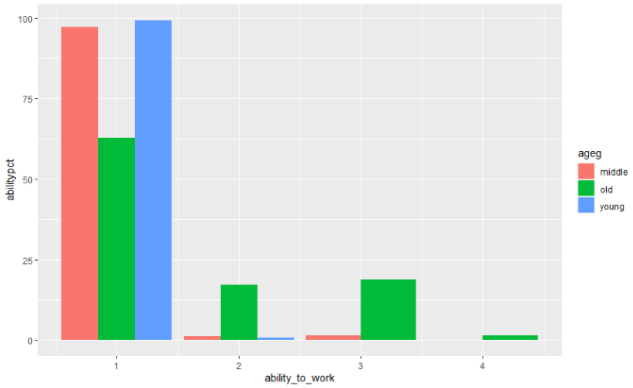
1번 미취학(만7세미만), 2번은 무학(만7세이상), 3번 초등학교, 4번 중학교, 5번 고등학교, 6번 전문대학, 7번 대학교, 8번 대학원(석사), 9 번대학원(박사) 이다. 위 그래프와 결과 값을 살펴보면 중, 고등학교 졸업은 여성이 전문대 이상의 교육 수준은 남성이 높은 것을 확인할 수 있다. 하지만, 그 비율은 비슷한 것으로 보아 남녀의 교육 수준은 크게 차이가 나지 않는 것을 확인할 수 있다.

(2) 나이에 따른 근로 능력 정도 (가구원 1)

나이에 따른 근로 능력을 비교해보고자 한다. 분석에 사용되는 변수는 근로 능력, 나이, 연령층이 있다. (1) 과 마찬가지로 노년의 나이가 가장 많은 분포를 갖고 있으므로 서로 정확한 파악이 어려워 근로 능력을 나눈 연령층에 따라 퍼센트를 활용하였다. 분석에 사용되는 변수들에서는 따로 결측치와 이상치가 확인이 되지 않아 is.na 함수를 사용하지 않고 분석하였다.

데이터 분석 후 결과 도출 시의 데이터에선, 청년, 중 장년, 노년층으로 나누어 1. 근로 가능, 2. 단순 근로 가능, 3. 단순 근로 미약자, 4. 근로 능력이 없어 경제적 능력을 하지 않음으로 나누어서 확인할 수 있다.

	ageg	ability_to_work	n	totalability	abilitypct
	<chr>	<dbl>	<int>	<int>	<dbl>
1	middle	1	1945	2003	97.1
2	middle	2	25	2003	1.2
3	middle	3	31	2003	1.5
4	middle	4	2	2003	0.1
5	old	1	2423	3857	62.8
6	old	2	661	3857	17.1
7	old	3	722	3857	18.7
8	old	4	51	3857	1.3
9	young	1	135	136	99.3
10	young	2	1	136	0.7



위 도출된 그래프를 확인해보면 4번에 있는 노년층의 나이는 고령(80세 이상)일 것으로 예상된다.

	ageg	ability_to_work	mean_age
	<chr>	<dbl>	<dbl>
1	old	4	82.4

위의 도출 예상과 같이 60대 이상이 노년 층이지만 노년층의 경계치인 60보다 22살 높은 평균 82.4가 근로능력이 없어 경제적 능력을 하지 않는 것으로 나타난다.

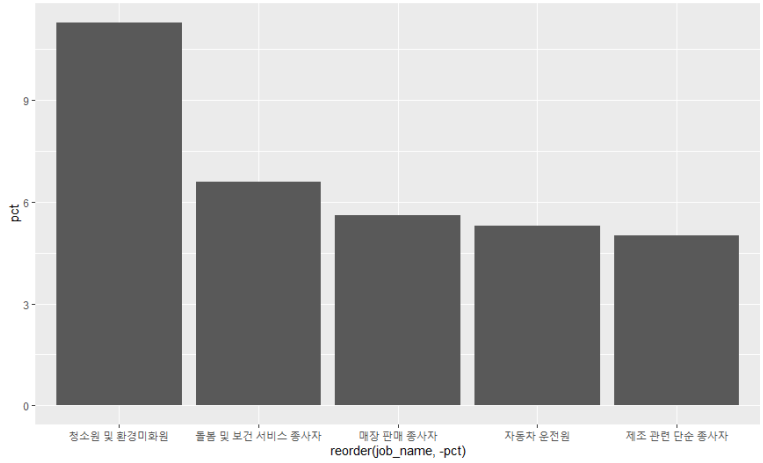
(3) 직업에 따른 이혼율 top 5(가구원1)

가구원 1의 직업에 따른 이혼율을 나타내었다. 데이터 분석 시, 직업이 없는 사람들은 배제하고, 직업이 있는 사람들을 중심으로 데이터 분석을 진행하였다.

2021년 16차 한국복지패널조사) 조사설계서-가구용(beta1) 데이터와 현재 주 데이터와 left join을 통해 연결을 하였다. Excel 파일의 4 시트에 있는 직종 데이터의 직업 이름(job_name)은 '..4'로 저장되어 있어 이름을 job_name으로 변경해주고, 코드도 code_job으로 변경해준다. join은 변수의 이름이 같을 때에만 가능하기 때문에 주 데이터에(data_file)도 변수의 이름을 변경해줘야 한다. 또한 엑셀에서 옮겨온 직업 데이터를 모두 사용하지 않아, dplyr select 함수를 통해 code_job, job_name만을 저장해준 후 join을 해준다.

join 후 데이터를 확인해보면, 직업이 없는 사람들은 NA로 결측치로 표현이 되어 있어 is.na 함수를 사용하여 데이터를 도출하였다.

	group_marriage	job_name	n	tot_group	pct
	<chr>	<chr>	<int>	<int>	<dbl>
1	divorce	청소원 및 환경미화원	36	319	11.3
2	divorce	돌봄 및 보건 서비스 종사자	21	319	6.6
3	divorce	매장 판매 종사자	18	319	5.6
4	divorce	자동차 운전원	17	319	5.3
5	divorce	제조 관련 단순 종사자	16	319	5



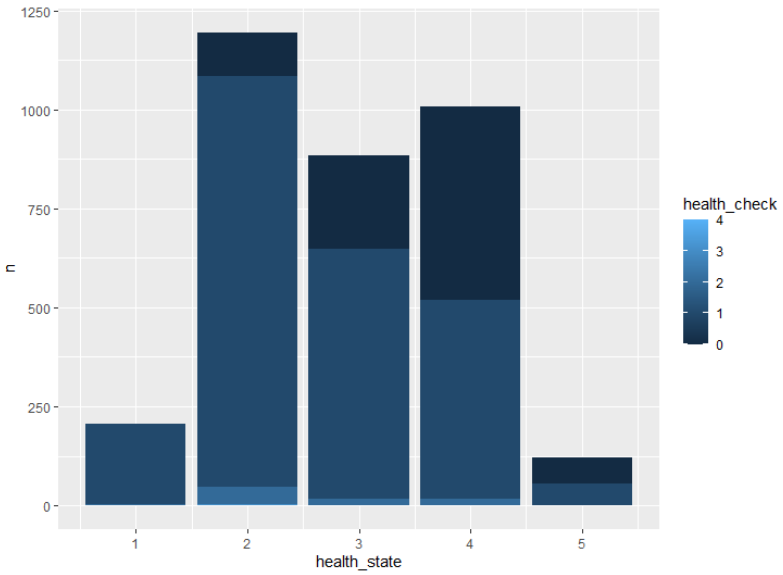
데이터 분석에 따르면 청소원, 돌봄 보건 서비스 종사자 매장 판매 종사자 순으로 이혼율이 높은 것을 확인할 수 있다. 하지만, 직업에 따른 이혼율이 완전히 높은 편은

아니며, 직종에서 가장 높은 청소원 및 환경 미화원은 10프로가 넘으며 나머지 데이터들은 6프로 미만인 것을 확인할 수 있다.

(4) 건강검진 횟수에 따른 건강 상태(가구원 1)

건강검진을 한 횟수에 따른 건강상태를 알아보고자 한다. 이 데이터 분석에서 필요한 변수는 건강상태, 건강검진 횟수가 있다. 전처리 과정에서 건강검진 데이터를 확인해보니 20번을 측정한 사람이 있어, 이상치로 확인되어 결측치(NA)로 만들어서 데이터 분석 시 배제하고 분석을 진행하였다. 건강 상태에서는 이상치와 결측치가 없었기에 is.na 함수 사용은 건강검진 횟수에서만 사용했다.

health_check	health_state	n
<dbl>	<dbl>	<int>
0	1	185
0	2	1195
0	3	883
0	4	1006
0	5	121
1	1	205
1	2	1084
1	3	649
1	4	519
1	5	56
2	1	3
2	2	48
2	3	17
2	4	16
2	5	1
3	1	1
3	2	4
3	3	1
4	4	1



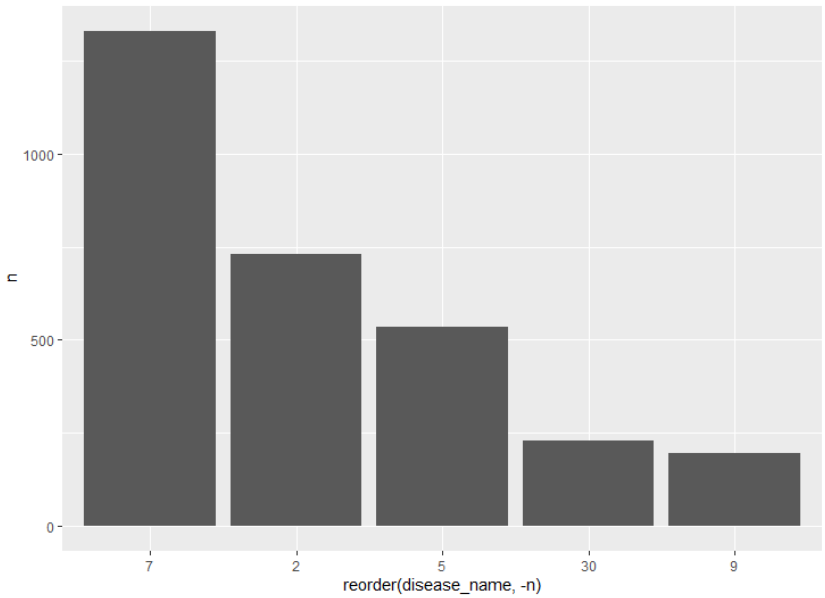
건강검진을 하지 않은 사람일수록 그래프에서 어두운 값으로 나타나는 것을 그래프

에서 확인할 수 있다. Health_state는 5로 갈수록 건강하지 않은 편에 속하는데, 건강검진 횟수가 적은 사람 일수록 건강 상태가 좋지 않은 것을 그래프에서 알 수 있다.

(5) 만성 질환 6개월 이상 투병중인 병명 중 높은 순위 top 5(가구원 1)

만성질환을 6개월 이상 투병중인 병명 중에 가장 높은 순위 top 5을 분석해 보고자 한다. 필요한 변수는 병명과 만성 질환 변수이다.

disease_name	n
<fct>	<int>
7	1330
2	729
5	534
30	230
9	195

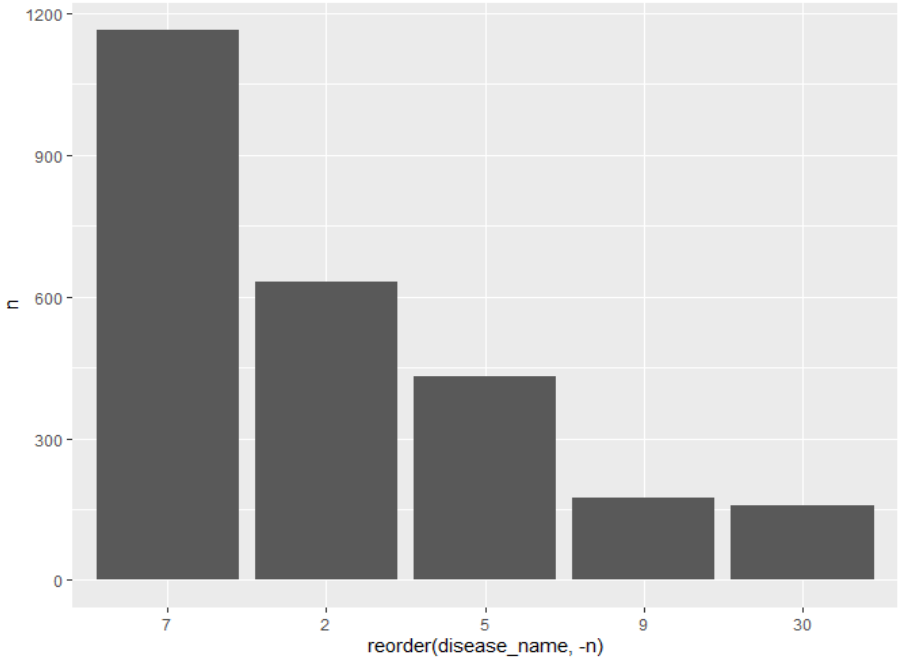


7은 고혈압, 2는 관절염, 요통, 좌골통, 디스크, 5는 당뇨병, 30은 기타질환(급성 질환 등) 9는 심근경색증, 협십증이다. 만성질환을 6개월 이상 투병중인 사람들 중에서 가장 높은 것은 고혈압이다.

그렇다면 노년 층에서 만성질환 6개월 이상 투병 중인 사람들도 이와 같은 그래프로 나타날까?

	disease_name	n
	<fct>	<int>
1	7	1330
2	2	729
3	5	534
4	30	230
5	9	195

```
> disease_6month_old<-
```



고위험군인 노년층에서는 7 고혈압, 2 관절염, 5당뇨병, 9 중풍, 뇌혈관 질환, 30 기타 질병 (급성 질환)으로 전 연령층과 비슷한 분포를 하고 있다. 하지만 4위와 5위가 기타질환 심근경색증이었던 전 연령층과 달리 노년층은 4위와 5위가 중풍, 뇌혈관 질환, 기타 질병인 것이 차이점이다. 노년에 들수록 뇌와 관련된 검사도 꾸준히 받아야 한다는 것을 데이터를 통해서 알 수 있다.