[비즈니스 모델링 프로젝트 과제]

● 서론

● 준비

- 분석 주제 선정
- 변수 선정
- 변수 전처리

● 관계 분석

- 연령대에 따른 건강상태 비율
- 연령대에 따른 근로 능력 정도 비율
- 연령대에 따른 종교 유무

● 마무리

● 서론

한국 보건사회 연구원에서는 복지 정책을 지원하기 위해 우리나라 사람들의 경제활동, 생활실태, 복지욕구 등을 조사해서 그 데이터를 한국 복지 패널이라는 사이트에 그 데이터를 남겨둔다. 이번 학기 비즈니스 모델링 수업에서는 이 사이트에 있는 위 데이터를 갖고 각 변수들을 분석하는 연습을 진행했는데, 이번 학기 프로젝트 과제가 바로 이 데이터에 있는 변수들을 통해 각 변수 간 관계를 분석하는 것이었다. 관계를 분석하는 과정은 분석 주제 선정, 변수 선정 및 전처리, 마지막으로 각 변수 간 관계를 분석하는 것이 있는데, 지금부터 위 과정이 프로젝트 과제에서 어떻게 사용되었는지 하나씩 알아보도록 하자.

● 준비

먼저 준비 단계이다. 이 단계에서는 분석 주제 선정부터 전처리까지 수행되는데, 이 과정을 통해 분석하기 위한 준비를 모두 끝내둔다. 준비 과정은 총 세 단계가 있는데, 먼저 분석 주제부터 선정해 보자.

- 분석 주제 선정

한국 복지 패널에서 제공 되는 데이터에는 수많은 변수가 있고, 이 변수들을 어떻게 조합하느냐에 따라 수많은 주제가 나올 수 있는데, 나는 수많은 주제들 중에 세 주제를 선정했다. 첫 번째는 연령대에 따른 건강상태 비율, 두 번째는 연령대에 따른 근로 능력 정도 비율, 세 번째는 연령대에 따른 종교 유무이다. 주제를 나눈 이유는 이렇다. 나는 이번 보고서를 통해 우리나라 사람들이 얼마나 건강한지, 그에 따른 근로 능력 정도가 어떻게 되는지, 종교는 갖고 있는지 알아보고 싶었는데, 나누는 기준은 다양하지만 나는 연령대에 따라 나누고자 하였고, 그에 따른 분석 결과가 어떻게 나오는지 알아보고자 하였다.

- 변수 선정

제공 데이터로부터 변수를 선정 해보자. 총 4개의 변수를 고를 수 있다. 첫 번째는 각 주제들에 공통으로 필요한 '연령대'인데, 연령대는 제공 데이터에 없기 때문에 대신에 '태어난 년도'로 선정했다. 그 나머지는 연령대에 따라 알고자 하는 변수들에 맞춰 각각 '건강 상태', '근로 능력 정도', '종교 유무'로 선정했다.

- 변수 전처리

변수는 선정했지만, 각 변수들은 원래 복지 패널이 사용하는 형식을 따르기 때문에 이들을 우리가 이해할 수 있도록 바꿔주는 과정이 필요하다. 이 과정이 바로 전처리 과정인데, 이 과정은 각 변수마다 적용되고, 전처리가 끝나면 비로소 변수 간 관계를 분석할 수 있다. 각 변수들의 전처리 과정에 대해 알아보자. 전처리 기준은 데이터 분석 매뉴얼로 쓰이는 한국 복지 패널 사이트의 '코드북'으로 두었다. 모든 전처리의 공통된 과정이 있다면 데이터에 기록된 이상치를 모두 결측치로 바꿔준다는 것이다. 이상치를 구분하는 기준은 변수마다 다르고, 그 기준 또한 '코드북'으로 두었다.

```r
# 태어난 년도 -> 연령대별로 전처리
savData$birth <- ifelse(savData$birth == 9999, NA, savData$birth)
savData$age <- 2021 - savData$birth + 1
savData <- savData %>%
  filter(!is.na(age)) %>%
  mutate(ageg = ifelse(age < 10, "0대",
                ifelse(age < 20, "10대",
                ifelse(age < 30, "20대",
                ifelse(age < 40, "30대",
                ifelse(age < 50, "40대",
                ifelse(age < 60, "50대",
                ifelse(age < 70, "60대",
                ifelse(age < 80, "70대",
                ifelse(age < 90, "80대", "90대 이상")))))))))
qplot(savData$ageg)
```

먼저 태어난 년도부터 알아보자. 각 데이터에서 기록된 값은 태어난 년도 4자리이고, 이상치는 9999인데, 이는 데이터를 모으기 위한 설문 조사 시 해당 문항의 무응답을 의미한다. 태어난 년도를 응답하지 않은 데이터에 대해서는 그 값을 결측치로 바꿨고, 그 이외의 값은 그대로 두었다. 그다음은 각 태어난 년도를 조사 년도에서 각 사람들의 태어난 년도를 빼고, 1을 더해서 한국식 나이를 파생 변수로 만들었고, 각 연령대를 10대부터 90대까지 나눠서 그에 따라 ageg라는 파생 변수를 만들었다. 만들어진 ageg가 바로 연령대에 해당되고, 세 주제 분석에서 다뤄지게 된다.

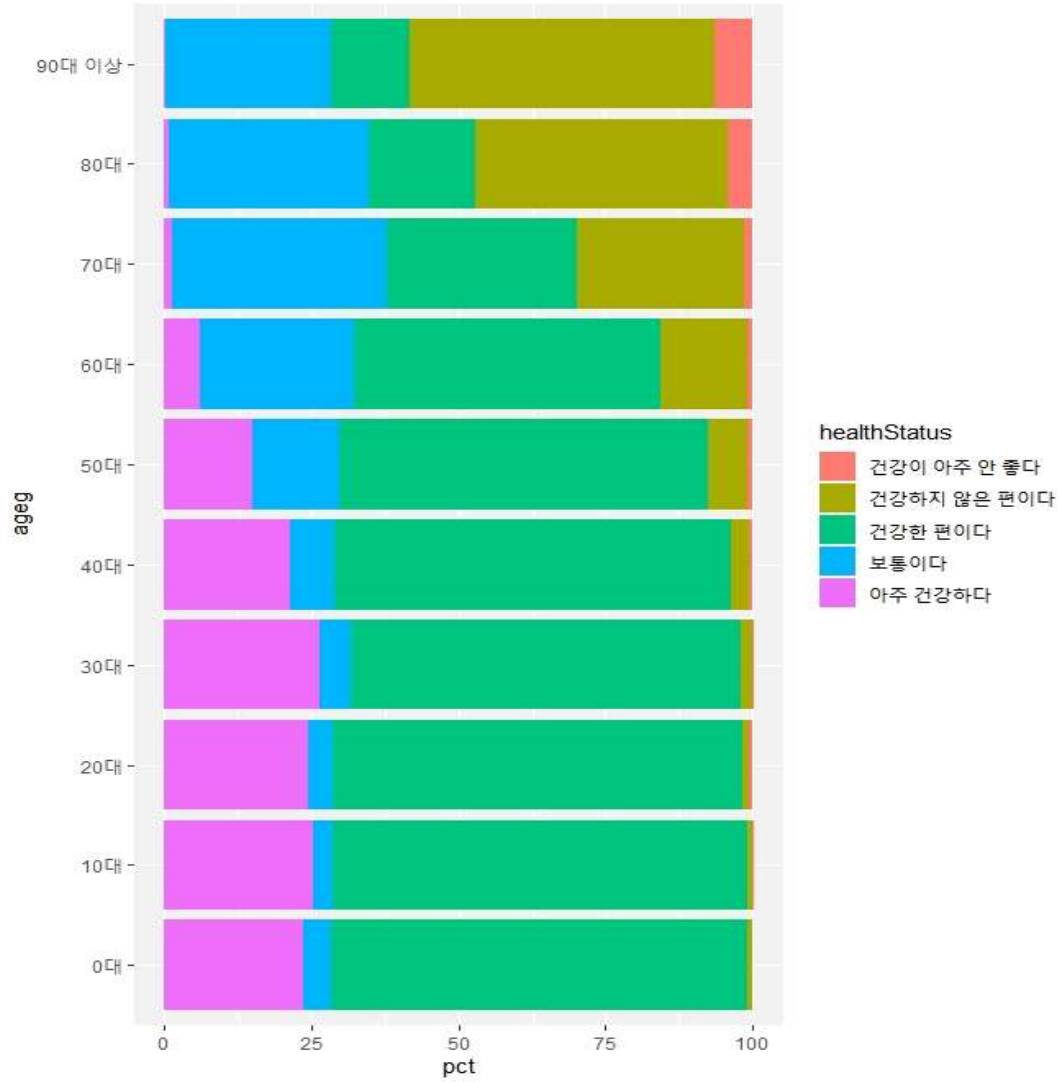
```r
# 건강 상태 전처리
savData$healthStatus <- ifelse(savData$healthStatus == 1, "아주 건강하다",
                        ifelse(savData$healthStatus == 2, "건강한 편이다",
                        ifelse(savData$healthStatus == 3, "보통이다",
                        ifelse(savData$healthStatus == 4, "건강하지 않은 편이다",
                        ifelse(savData$healthStatus == 5, "건강이 아주 안 좋다", NA)))))
qplot(savData$healthStatus) + coord_flip()
# 근로 능력 전처리
savData$workAbility <- ifelse(savData$workAbility == 1, "근로 가능",
                       ifelse(savData$workAbility == 2, "단순 근로 가능",
                       ifelse(savData$workAbility == 3, "단순 근로 미약자",
                       ifelse(savData$workAbility == 4, "근로 능력 없음", NA))))
qplot(savData$workAbility) + coord_flip()
# 종교 유무 전처리
savData$region <- ifelse(savData$region == 1, "있음", "없음")
qplot(savData$region)
```

다음은 건강 상태, 근로 능력, 종교 유무이다. 건강 상태의 경우 데이터에 기록된 값은 1부터 5까지이고, 각 값들은 각각 '아주 건강하다'부터 '건강이 아주 안 좋다'까지 다섯 가지의 내용에 해당된다. 근로 능력의 경우는 1부터 4까지이고, 각 값들은 '근로 가능'부터 '근로 능력 없음'까지 네 가지의 내용에 해당된다. 종교 유무의 경우, 1이면 있음, 2면 없음에 해당되는데, 결측치가 없는 종교 유무를 제외하고, 나머지 두 변수에서 범위 내 값이 없다면 결측치로 처리했다.

● 관계 분석

이제 전처리 과정을 거친 변수들을 통해 선정한 주제에 대한 변수 간 관계를 분석해 보자. 분석한 결과는 각 주제마다 나올 그래프로 설명되고, 해당 그래프를 통해 각 변수가 선정한 주제에 따라 어떠한 관계를 갖고 있는지 설명될 수 있다. 그럼 먼저 연령대에 따른 건강 상태부터 알아보자.

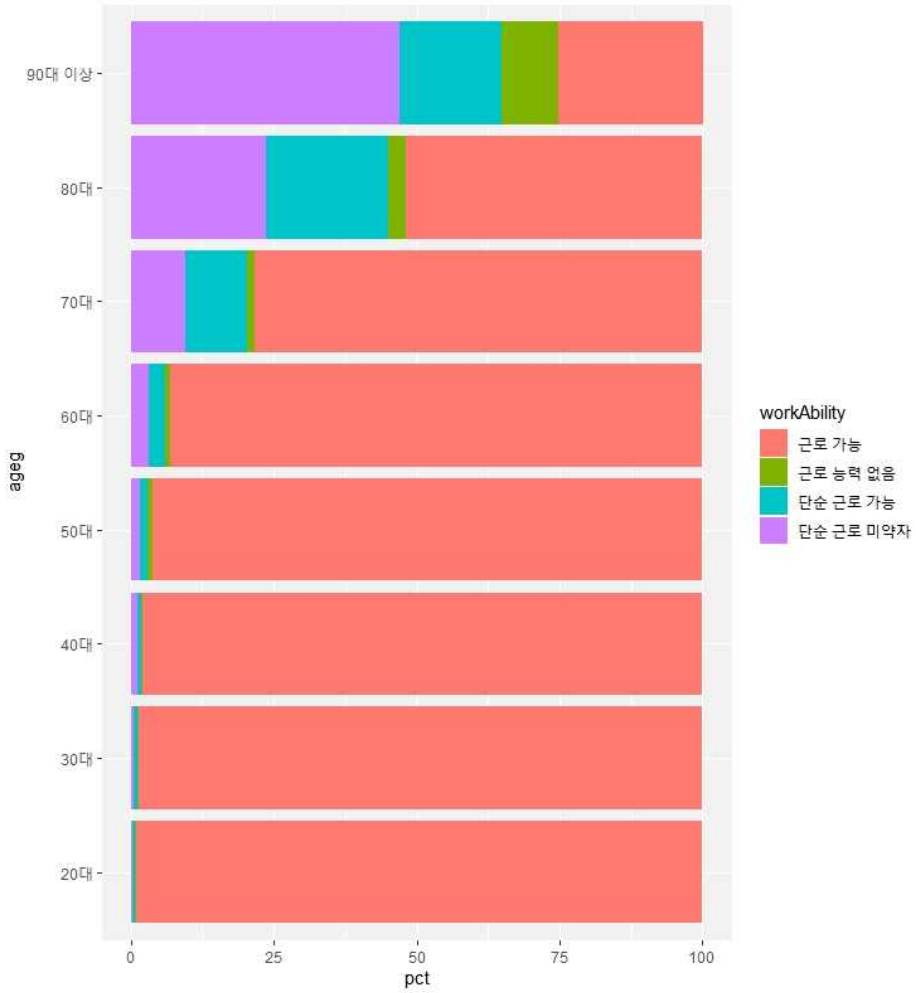
- 연령대에 따른 건강상태 비율




```
# 1. 연령대에 따른 건강 상태 비율
ageg_healthStatus <- savData %>%
                     group_by(ageg, healthStatus) %>%
                     summarise(statusNumber = n()) %>%
                     mutate(tot_group = sum(statusNumber)) %>%
                     mutate(pct = round(statusNumber/tot_group*100, 1))
ageg_healthStatus
ggplot(data = ageg_healthStatus, aes(x = ageg, y = pct, fill = healthStatus))
+ geom_col() + coord_flip()
```

위 그래프는 연령대에 따른 건강 상태 비율을 나타낸다. 먼저 전처리가 완료된 데이터에서 연령대와 건강 상태를 기준으로 그룹을 묶었고, 그룹 내 각 행 수를 계산했으며, 파생 변수로 각 행 수의 합, 그룹 내 건강 상태 비율을 계산했다. 분석 결과 나이가 아주 건강하다는 결과는 30대까지 일정하다가 40대부터 줄어서 80대 이후에는 그 수치가 거의 안 보이고, 건강하다는 40대까지는 일정하다가 50대부터 점점 줄어들었으며, 보통이다는 0대부터 70대까지 계속 증가하였고, 건강하지 않은 편과 건강이 아주 안 좋은 경우는 40대부터 90대까지 계속 증가하는 것을 알 수 있었다. 나이가 들면서 사람의 건강이 점차 안 좋아지는 것은 상식으로 알고 있지만, 위 그래프에서 그 사실이 통계적으로 드러나는 것을 알 수 있다.

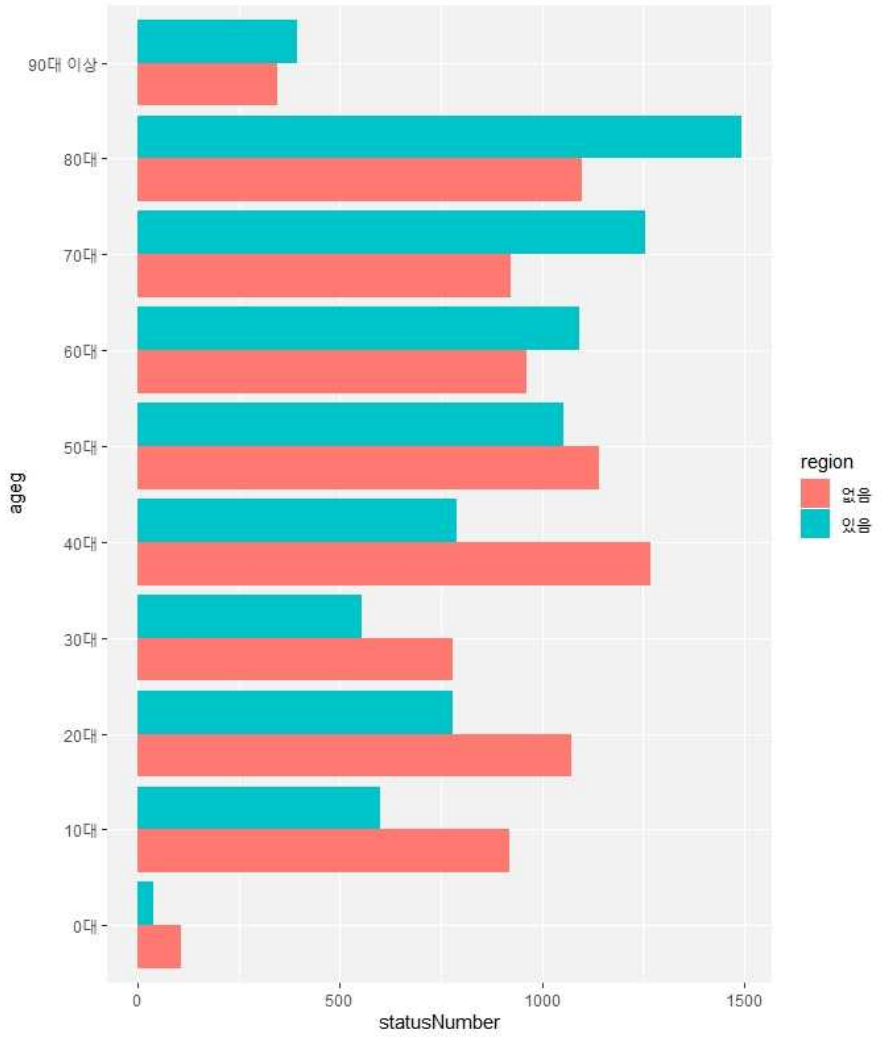
- 연령대에 따른 근로 능력 정도 비율




```
# 2. 연령대에 따른 근로 능력 정도
ageg_workAbility <- savData %>%
                    filter(!is.na(workAbility)) %>%
                    group_by(ageg, workAbility) %>%
                    summarise(statusNumber = n()) %>%
                    mutate(tot_group = sum(statusNumber)) %>%
                    mutate(pct = round(statusNumber/tot_group*100, 1))
ageg_workAbility
ggplot(data = ageg_workAbility, aes(x = ageg, y = pct, fill = workAbility)) +
  geom_col() + coord_flip()
```

다음은 연령대에 따른 근로 능력 정도 비율이다. 먼저 근로 능력이 결측치인 값은 없앴고, 연령대와 근로 능력으로 그룹을 묶은 다음, 앞선 경우와 마찬가지로 그룹에 포함된 각 행의 수를 셌다. 이후 행수의 합을 구했고, 이를 통해 비율을 구했는데, 나온 결과는 ageg_workAbility에 입력했다. 입력 결과로 그래프를 그린 결과는 위와 같은데, 근로 가능 여부는 50대부터 서서히 줄어들다가 70대부터 급격하게 줄어드는 것을 알 수 있다. 또한 단순 근로 가능 여부는 오히려 나이가 들수록 증가하는 것을 알 수 있고, 단순 근로 미약자와 근로 능력 없음 또한 나이가 들면서 증가하는데, 두 수치는 모두 90대 이상에서 제일 크게 나타나는 것을 알 수 있다. 건강 상태와 마찬가지로 근로 능력 또한 나이가 들면서 일을 할 수 있는 사람이 점점 줄어드는 것을 통계적으로 알 수 있다.

- 연령대에 따른 종교 유무




```
# 3. 연령대에 따른 종교 유무
ageg_region <- savData %>%
                group_by(ageg, region) %>%
                summarise(statusNumber = n()) %>%
                mutate(tot_group = sum(statusNumber)) %>%
                mutate(pct = round(statusNumber/tot_group*100, 1))
ageg_region
ggplot(data = ageg_region, aes(x = ageg, y = statusNumber, fill = region))
+ geom_col(position = "dodge") + coord_flip()
```

마지막은 연령대에 따른 종교 유무이다. 앞선 두 경우와 다르게 이는 비율이 아닌, 조사된 사람의 수로 계산했고, 각 막대를 종교가 있는 경우와 없는 경우를 따로 표시했다. 데이터를 알아내는 과정은 두 경우와 같다. 연령대와 종교를 그룹으로 묶었고, 그룹에 속한 행의 수를 셌으며, 행의 수 합산과 비율까지 구하긴 했지만, 위 그래프에서는 비율을 사용하진 않았고, 각 사람의 수를 나타내는 statusNumber 변수를 사용하였다. 그래프를 분석해 보자. 먼저 그래프의 추이를 보면 10대부터 40대까지 종교가 없는 사람은 점점 늘어나고, 50대부터 70대 까지는 종교가 없는 사람이 점점 줄어들며, 80대에 잠깐 종교가 없는 사람이 늘어난 것을 알 수 있다. 종교가 있는 경우는 10대부터 30대 까지는 20대 기준으로 늘었다가 줄어들지만, 이후 80대 까지는 종교를 갖는 사람이 계속 늘어남을 알 수 있다. 나이가 들면서 건강 상태가 안 좋아지거나 근로 능력이 떨어지는 변화를 보이지만, 종교를 믿는 사람은 반대로 늘어나는 것을 알 수 있다.

● 마무리

지금까지 세 주제 분석해 보았다. 분석하면서 느낀 점은 분석보다 중요한 것은 역시 변수 선정 및 전처리라는 것이다. 솔직히 분석하는 과정은 그리 길지 않았고, 각 변수를 선정하고, 분석하는 사람이 이해할 수 있게 전처리하는 과정이 더 길었던 것이 사실이다. 올바른 변수를 선정하지 않거나, 전처리를 제대로 거치지 않았다면 위처럼 우리가 원하는 결과를 얻지 못했을 것이다. 우리가 정한 주제에 대한 정확한 분석 결과를 얻기 위해 분석에 필요한 코드를 잘 구성하는 것 또한 매우 중요하지만, 그보다 중요한 것이 바로 분석을 위한 준비과정임을 느끼게 된다. 데이터를 분석하면서 느낀 점은 우리가 상식으로 알고 있는 부분이 통계적으로 표현될 때 어떻게 표현될 수 있는지, 어떤 코드를 작성하면 되는지 잘 알게 되었다는 점이다. 사실 이 수업은 통계 수업이 아닌 데이터마이닝 수업인 것은 맞다. 통계는 가설을 두고 그 사실이 맞는지 통계적으로 검증해 내지만, 데이터마이닝은 가설을 두지 않고, 정한 주제에 따른 분석 결과가 어떻게 되는지 도출해 낸다. 이 사실을 분명 알고 있지만, 상식적으로 연령대에 따라 건강 상태가 어떤지, 근로 능력 정도가 어떤지는 모르는 사람이 없을 것이다. 그럼에도 불구하고 이 주제를 선정한 이유는 우리가 이 상식을 모른다고 가정했을 때 어떻게 이러한 상식이 만들어졌는지 통계적인 기법으로 도출되었는지 알아내기 위해서이다. 우리가 아는 상식은 분명 무언가에 근거해서 만들어졌을 것인데, 근거가 되는 것을 알아내기 전에는 연령대가 많아지면서 건강 상태가 어떻게 되는지, 근로 능력 정도가 어떻게 되는지 알 수 없었을 것이다. 결국 알 수 없었던 것에서 근거를 만들어내고, 그 근거를 바탕으로 우리의 상식이 만들어졌을 것이기에, 우리가 상식을 몰랐다고 가정하고 데이터를 분석하면 이 또한 데이터마이닝이라고 할 수 있다. 사실은 이 외에도 분석할 수 있는 변수의 관계는 충분히 많다. 그래도 나는 우리가 쉽게 알 수 있는 변수로 선정하고자 하였고, 그 변수를 통해 무언가 결과를 도출해 내는데 성공하였다. 앞으로 데이터의 변수를 분석할 수 있는 기회가 생긴다면 상식을 모른다고 가정하는 것이 아닌 상식적으로 정말 모르는 변수 간의 관계에 대해 분석해 보고 싶다. 예를 들면 근로 능력에 따른 종교 유무처럼 전혀 연관이 없고, 상식적으로 관련이 없어 보이는 두 변수 사이에도 데이터마이닝 분석 기법을 적용하면 변수 간 연관성을 발견할 수도 있는 것이다. 그러한 식으로 상식을 만들어 나가고, 만들어진 상식은 미래의 사회, 경제, 비즈니스의 발전 과정을 예측할 수 있는 근간이 될 것이라고 난 믿는다.

[참고 문헌]

데이터 및 코드북 참고 : http://www.koweps.re.kr:442/data/data/list.do