|  |
| --- |
| **빅데이터R분석실습 – 복지 데이터 분석****한국 복지 패널 데이터로 저소득 가구와 지역간의 관계 조사** |

****

|  |  |  |  |
| --- | --- | --- | --- |
| **제출일** | **2019-12-15** | **전공** | **컴퓨터공학과** |
| **과목** | **빅데이터R분석실습** | **학번** | **20184060** |
| **담당교수** | **정치봉 교수님** | **이름** | **안성현** |

< 목 차 >

01 데이터 조사

1-1] 데이터 세트 R-studio로 가져오기

1-2] ‘한국 복지 패널 데이터’에 관해서 조사하기

02 전처리 과정

2-1] 파일에서 필요한 변수 가져오기

2-2] 데이터 가공하기

2-3] 데이터 정제하기

03 데이터 분석

3-1] 전처리된 데이터 기초적인 탐색하기

3-2] ‘지역 별 저소득 가구의 인구 수’ 시각화 및 분석

3-3] ‘지역 별 저소득 가구의 가구원 수’ 시각화 및 분석

3-4] ‘지역 별 저소득 가구의 가구형태’ 시각화 및 분석

3-5] ‘지역 별 저소득 가구의 소득’ 시각화 및 분석

3-6] ‘지역 별 저소득 가구의 나이대’ 시각화 및 분석

3-7] ‘지역 별 저소득 가구의 인구 구분’ 시각화 및 분석

04 부록(전체 코드)

 데이터 조사

1-1] 데이터 세트 R-studio로 가져오기

**① 코드**

|  |
| --- |
| # (1) 분석 작업을 위해 자신의 작업 환경 및 데이터 준비# 및 한국복지패널 데이터 R-studio로 불러오기library(foreign)f<-read.spss(file="Koweps\_hpc10\_2015\_beta1.sav",to.data.frame=T)# (2) 한국복지패널 데이터(f)의 기초적인 탐색하기str(f) # 데이터 세트의 구성colnames(f) # 변수의 이름dim(f) # 행과 열의 개수head(f,3) # 일부 데이터만 보기 View(f) # 전체 데이터 보기summary(f) # 요약본 보기 |

**② 결과물**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ▶ str함수를 이용해서 데이터 세트의 구성을 확인했다. 16664개의 행과 957개의 열을 가지므로 약 1600만개의 데이터를 가지는 파일이다. 각 변수는 수치형 데이터를 가지고 있다.▶ 데이터 전체를 보기 전에 summary함수로 데이터 세트를 요약해봤다. 각 변수들의 최소값이 출력된다. 변수가 너무 많고 변수의 이름도 복잡하므로 요약 데이터로는 이해하기 어렵다.▶ view함수를 이용해서 데이터 전체를 확인하였다. 각 행마다 한 사람의 복지 데이터를 가진다. (1행은 a사람 데이터, 2행은 b사람 데이터, …) 각 변수 밑에는 이 변수가 어떤 데이터인지 한국말로 기재돼있다. 이 기재된 내용이 ‘한국복지패널데이터’를 이해하는데 많은 도움을 주었다.1-2] ‘한국 복지 패널 데이터’에 관해서 조사하기 한국보건사회연구원에서 가구의 경제 활동 및 상태를 연구하여 정책지원에 반영할 목적으로 조사한 자료이다.   2006~2015년까지 전국 7000여 가구를 선정해 매년 추적 조사한 자료로, 경제활동, 생활실태, 복지 요구등 약 1000개의 변수를 가진 데이터 세트이다.그러나 필자가 조사할 것은 ‘저소득 가구와 지역간의 관계’이므로 모든 변수를 사용하지 않는다. 아래에 설명한 변수는 분석할 때 필요한 변수만을 이용했다.h10\_id : 조사 대상 가구에 부여된 IDh10\_reg7 : 7개 권역별 지역 구분<1.서울, 2. 수도권(경기,인천), 3. 부산/경남/울산, 4. 대구/경북, 5. 대전/충남, 6. 강원/충북 7. 광주/전북/전남/제주>h10\_din : 가처분소득 (한 해 소득에서 세금 및 이자를 제외한 값)h10\_hc : 소득에 따른 가구 구분 <1. 일반 가구, 2. 저소득층 가구>h1001\_1 : 가구원 수h10\_g2 : 가구주와의 관계<10. 가구주, 20. 가구주의 배우자, 11~16. 가구주의 자녀, 111~166. 가구주의 손주, 21~26: 가구주 자녀의 배우자>C:\Users\안성현\Desktop\관계.jpgh10\_g3: 성별h10\_g4 : 태어난 연도h10\_110 : 가구 형태<1.단독, 2.모자, 3.부자, 4. 조손 가구 or 소년/소녀 가장, 5. 기타(부모와 자식 같이 생활, 결혼 후 부모님 모시고 생활, 동거인과 생활 등)>  전처리 과정 2-1] 파일에서 필요한 변수 가져오기**① 코드**

|  |
| --- |
| # 필요한 변수 정하기 default<-"h10"vector<-c("\_id","\_reg7","\_din","\_hc","01\_1","\_g2","\_g3","\_g4","01\_110")vector\_name<-c("가구번호","지역","가처분소득","가구구분","가구원수","가구주와의관계","성별","태어난연도","가구형태")# 필요한 변수만 가져와서 리스트 만들기file<-list()for(i in 1:9){ file<-c(file,f[paste(default,vector[i],sep='')])}# 리스트를 데이터 프레임으로 변환하고 열 이름 바꾸기 file<-as.data.frame(file)for(i in 1:9){ colnames(file)[i]<-vector\_name[i]}# 필요한 변수만 가져온 데이터 보기 View(file) |

**② 결과물**

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ▶ subset이나 select함수를 쓰는 방법도 있겠지만 필자는 paste함수를 써서 필요한 변수를 가져왔다. h10\_id, h10\_reg7, h10\_din 등의 변수에는 ‘h10’ 이라는 문자열이 공통적으로 쓰인다. 그래서 default변수에는 ‘h10’을 저장하고 vector벡터에는 ‘\_id,\_reg7,\_din, …’ 을 저장한 뒤 paste함수로 default와 vector의 각 요소를 합쳤다. 합치는 과정은 for문에서 i가 1씩 증가되면서 default와 vector의 i번째 요소를 결합하는 방식이다. 이후 f(데이터세트)에서 결합된 문자열을 변수로 가지는 데이터들을 불러왔다. 이 과정을 i가 1부터 9까지 진행한다. 진행하면서 불러온 데이터들은 합쳐져서 file이라는 리스트에 저장된다. file리스트는 데이터 프레임 형태로 전환된다. 그리고 각 변수의 이름을 바꾸는 작업이 시작된다. 이 작업을 위해 vector\_name이라는 벡터를 미리 만들었다. h10\_id는 가구 ID이니 ‘가구번호’로, h10\_reg7은 7개 권역별 지역 구분이니 ‘지역’으로 하며 총 9개의 변수 이름을 생각해서 vector\_name벡터에 저장했다.변수의 이름을 바꾸는 방법은 colnames(file)의 i번째 요소를 다른 문자열로 바꾸면 된다. 이 과정 역시 for문으로 진행됐다.2-2] 데이터 가공하기**① 코드**

|  |
| --- |
| # file 데이터 세트 가공하기 # 각 행의 데이터를 문자열로 보이게 하기file$"지역"<-factor(file$"지역",levels=c(1,2,3,4,5,6,7),labels=c("서울","수도권","경남","경북","충남","강원/충북","전라/제주"))file$"성별"<-factor(file$"성별",levels=c(1,2),labels=c("남","여"))file$"가구구분"<-factor(file$"가구구분",levels=c(1,2),labels=c("일반 가구","저소득층 가구"))file$"가구형태"<-factor(file$"가구형태",levels=c(1,2,3,4,5),labels=c("단독","모자/모녀","부자/부녀","조손가구","기타"))# "가구주와의관계"가 너무 복합함 table(file$"가구주와의관계")# "가구주와의관계" 데이터를 (함축해서) 문자열로 보이게 하기file$"가구주와의관계"<-ifelse(11<=file$"가구주와의관계" & file$"가구주와의관계"<=16,"가구주의 자녀", ifelse(111<=file$"가구주와의관계" & file$"가구주와의관계"<=166,"가구주의 손주", ifelse(21<=file$"가구주와의관계" & file$"가구주와의관계"<=26,"가구주 자녀의 배우자", ifelse(file$"가구주와의관계"==10,"가구주", ifelse(file$"가구주와의관계"==20,"가구주의 배우자",file$"가구주와의관계")))))file$"가구주와의관계"<-as.factor(file$"가구주와의관계") |

**② 결과물**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ▶ 위는 file(필요한 변수만 가져와서 만든 데이터 프레임)을 가공한 결과이다. 지역, 가구 구분, 가구주와의 관계, 성별, 가구형태가 수치형에서 문자열로 변한게 핵심이다.이렇게 데이터의 형태를 변환할 때는 factor함수를 이용하는게 최선이다. file$"지역"<-factor(file$"지역",levels=c(1,2,3,4,5,6,7),labels=c("서울","수도권","경남","경북","충남","강원/충북","전라/제주")) 소스를 확인하자.이는 file프레임의 ‘지역’변수를 가지고 factor들을 만드는데, ‘지역’변수에서 데이터 1~7을 을 ‘서울’,’수도권’,’경남’,…으로 비꿔서 만드는 것을 의미한다.다른 변수도 이와 같은 과정으로 데이터의 형태를 변화시켰다.다만 ‘가구주와의 관계’의 데이터는 아래와 같이 많기에 factor함수로 만들기에 무리가 있었다.고로 11~16은 ‘가구주의 자녀’로 111~166은 ‘가구주의 손주’로 21~26 ‘가구주 자녀의 배우자’로 비슷한 데이터를 한 번에 묶는 방법을 취하였다.이 과정은 소스의 ifelse문을 확인하면 된다.2-3] 데이터 정제하기1. **코드**

|  |
| --- |
| # 가구형태가 "기타"인 데이터가 너무 복잡하므로 삭제 # (부모,자녀가 같이 사는 가구, 룸메이트끼리 사는 가구, 친인척이랑 사는 가구 등) 삭제 table(file$"가구형태")file<-filter(file,file$"가구형태" != "기타")# 가구구분이 "저소득가구" 데이터만 남겨놓음table(file$"가구구분")file<-filter(file,file$"가구구분" != "일반 가구")# "저소득 가구"데이터 밖에 없으니 "가구구분" 변수 삭제file<-select(file,-가구구분)# 데이터 정제 (이상치 처리)file<-filter(file,file$가처분소득>0) # 소득이 마이너스인 가구 제거 file$"가구주와의관계"<-drop.levels(file$"가구주와의관계") # 값이 0인 레벨 제거file$"가구형태"<-drop.levels(file$"가구형태") # 값이 0인 레벨 제거file$"가구구분"<-drop.levels(file$"가구구분") # 값이 0인 레벨 제거 |

1. **결과물**

|  |  |
| --- | --- |
| **▶** 위는 file(가공된 데이터 세트)을 정제한 결과이다. 데이터 정제는 이상한 데이터를 처리하거나 삭제하는 것을 말한다. 첫 번째 바뀐 점은 ‘가구 형태’변수에서 데이터가 ‘기타’인 행을 삭제한 점이다. 가구 형태가 명확하지 않고 경우의 수가 너무 많기 때문에 이를 포함하면 데이터 분석 중 문제가 발생할 수 있을 것이라 판단했다.두 번째 바뀐 점은 ‘가구 구분’변수에서 데이터가 ‘일반 가구’인 행을 삭제한 점이다.필자의 분석은 저소득층 가구에 대한 것이므로 일반 가구 데이터는 사용할 필요가 없다. 고로 ‘저소득층’데이터만 남겨 두고 ‘가구 구분’ 변수도 삭제했다.이는 언뜻보면 이상한 데이터가 아니지만 주제에 맞지 않는 데이터이기에 이상한 데이터라고 생각하고 정제한 것이다. 위 두 과정은 dplyr패키지의 filter와 select로 쉽게 할 수 있다.세 번째 바뀐 점은 소득이 마이너스인 행을 삭제한 점이다.이는 주제와 상관없이 명확하게 이상하므로 정제할 필요가 있다. 지금까지 과정이 마친 후 drop.level함수로 삭제하고 남은 데이터를 정리하였다. |  |

 데이터 분석3-1] 전처리된 데이터 기초적인 탐색하기**① 코드**

|  |
| --- |
| # (6) 가공된 한국복지패널데이터(file) 의 기초적인 탐색하기str(file) # 데이터 세트의 구성 colnames(file) # 변수의 이름dim(file) # 행과 열의 개수head(file,3) # 일부 데이터만 보기 View(file) # 전체 데이터 보기summary(file) # 요약본 보기 |

1. **결과물**

**▶** 위는 전처리된 데이터 세트의 구성을 본 것이다. 8개의 변수와 1729개의 행을 가진다. 변수는 (가구번호,지역,가처분소득,가구원수,가구주와의관계,성별,태어난연도,가구형태)를 가지며 숫자형이거나 팩터형인게 특징이다. **▶** 위는 전처리된 데이터 세트의 요약본을 본 것이다. 가구번호는 1번부터 9797번까지 존재한다. 서울 사는 저소득층 사람은 186명, 수도권은 281명, 경남은 341명 등이다. 저소득층 사람의 가처분 소득 최저는 43만원이고 최고는 3279만원이다.가구원수의 최저는 1명, 최대는 5명이다. 이외에도 요약된 데이터는 위 이미지를 보고 확인할 수 있다!**▶** View함수로 전체 데이터를 확인하면 위 이미지와 같다.3-2] ‘지역 별 저소득 가구의 인구 수’ 시각화 및 분석**① 코드**

|  |
| --- |
| # (7) 데이터 분석: 탐색, 시각화, 분석# 데이터를 편하게 사용하기 위해 attach 함수를 사용attach(file)# 지역 별 저소득 가구의 인구 수people\_count<-table(지역)pie(people\_count,xlab="지역 별 저소득 가구의 인구 수")# 지역 별 저소득 가구의 인구 수 (성별 포함)people\_count2<-table(성별,지역)barplot(people\_count2,xlab="지역 별 저소득 가구의 인구 수",col=c("red","darkblue"),beside=T)legend("topleft",levels(성별),lty=c(1,1),col=c("red","blue")) |

1. **결과물**

**▶ ‘**지역별 저소득 가구의 인구 수’ 이다. ‘전라/제주’가 저소득 인구가 가장 많고 다음이 경남 그 다음이 수도권이다. 이 데이터는 summarise를 이용한 요약본에도 나와 있기 때문에 성별도 추가해서 다시 시각화했다. 그렇게 했을 때 여자가 남자보다 압도적으로 저소득 가구에 속해 있었다. 또한 수도권보다 경북에 여성 저소득층이 많다는 사실을 알 수 있었다. 3-3] ‘지역 별 저소득 가구의 가구원 수’ 시각화 및 분석**① 코드**

|  |
| --- |
| # 지역 별 저소득 가구의 가구원 수people\_fcount<-table(가구원수,지역)barplot(people\_fcount,xlab="지역 별 저소득 가구의 가구원 수",col=c("red","orange","yellow","green","blue"),beside=T)legend("topleft",levels(as.factor(가구원수)),lty=c(1,1,1,1,1),col=c("red","orange","yellow","green","blue")) |

1. **결과물**

**▶** 가구원 수는 1명~5명까지 존재한다. 1명만 사는 저소득 가구는 전라/제주>경남>경북 순으로 많다. 2명이 사는 저소득 가구는 전라/제주>수도권>경북순으로 많다. 위의 시각화 자료를 보고 2명이서 사는 저소득 가구는 경상도보다 수도권에 생활하는게 나을 수 있다는 추론이 가능하다. 3-4] ‘지역 별 저소득 가구의 가구형태’ 시각화 및 분석**① 코드**

|  |
| --- |
| # 지역 별 저소득 가구의 가구형태 people\_type<-table(가구형태,지역)barplot(people\_type,xlab="지역 별 저소득 가구의 가구형태",col=c("red","orange","yellow","green"),beside=F)legend("topleft",levels(가구형태),lty=c(1,1,1,1),col=c("red","orange","yellow","green")) |

1. **결과물**

**▶** 이번에는 가구 형태를 알아 보았다. 가구 형태는 (단독,모자/모녀,부자/부녀,조손가구)가 존재한다. 이 데이터를 보면 전라/제주가 저소득 인구는 가장 많은데, 조손 가구는 경남이랑 비슷한 수치를 보인다. 또한 모자/모녀 저소득 가구는 수도권이 압도적으로 많은 것을 알 수 있다. 그리고 부자/부녀 저소득 가구는 모든 지역이 거의 없지만 경북에서는 압도적으로 없는 것을 확인할 수 있다.  |  |

 |  |

3-5] ‘지역 별 저소득 가구의 소득’ 시각화 및 분석**① 코드**

|  |
| --- |
| # 지역 별 저소득 가구의 소득 people\_money<-as.data.table(summarise(group\_by(file, 지역), mean(가처분소득)))colnames(people\_money)<-c("지역","가처분소득\_평균")ggplot(people\_money, aes(지역,가처분소득\_평균))+geom\_bar(stat="identity",fill="green") |

1. **결과물**

**▶** 지역마다 저소득 가구의 소득 평균은 얼마나 차이날지 ggplot을 이용해서 한 번 알아보았다. 인구는 전라/제주가 가장 많지만 소득은 수도권이 가장 많았다. 그다음은 경북>충남>서울 순이다. 참고로 수도권의 저소득 가구 인구 수는 3위이다. 경북은 수도권과 거의 차이 안 나게 4위이지만 소득 평균은 가장 낮은 것을 확인할 수 있다. 인구 수: 전라/제주 > 경남 > 수도권 > 경북 > 서울 > 충남 > 강원/충북 소득 평균: 수도권 > 경남 > 충남 > 서울 > 강원/충북 > 전라/제주 > 경북3-6] ‘지역 별 저소득 가구의 나이대’ 시각화 및 분석**① 코드**

|  |
| --- |
| # 지역 별 저소득 가구의 나이대 file$나이대<-ifelse(태어난연도<=1950,"노년층", ifelse(태어난연도<=1970,"중년층", ifelse(태어난연도<=1990,"장년층", ifelse(태어난연도<=2005,"청년층","아동층"))))people\_age<-table(file$나이대,지역)barplot(people\_age,xlab="지역 별 저소득 가구의 나이대",col=c("red","orange","yellow","green","blue"),beside=T,horiz=T)legend("topright",levels(as.factor(file$나이대)),lty=c(1,1,1,1,1),col=c("red","orange","yellow","green","blue")) |

1. **결과물**

**▶** 이번엔 지역과 저소득 가구의 나이대를 한 번 관련지어봤다. 전체적으로 노년층이 저소득 가구에 가장 많이 속한다. 그중에서 전라/제주가 가장 수치가 높고 다음은 경남>경북이다.젊은 층인 장년층 같은 경우 수도권에서 가장 많은 것을 알 수 있다. 다 같은 저소득 가구여도 지역마다 나이대 차이는 크다는 것을 알 수 있다. 3-7] ‘지역 별 저소득 가구의 인구 구분’ 시각화 및 분석**① 코드**

|  |
| --- |
| # 지역 별 저소득 가구의 인구 구분 서울<-subset(file,file$지역=="서울")서울$인구구분<-ifelse(서울$가구주와의관계=="가구주","서울\_가구주", ifelse(서울$가구주와의관계=="가구주의 배우자","서울\_가구주의 배우자", ifelse(서울$가구주와의관계=="가구주의 손주","서울\_가구주의 손주", ifelse(서울$가구주와의관계=="가구주의 자녀","서울\_가구주의 자녀",서울$가구주와의관계))))수도권<-subset(file,file$지역=="수도권")수도권$인구구분<-ifelse(수도권$가구주와의관계=="가구주","수도권\_가구주", ifelse(수도권$가구주와의관계=="가구주의 배우자","수도권\_가구주의 배우자", ifelse(수도권$가구주와의관계=="가구주의 손주","수도권\_가구주의 손주", ifelse(수도권$가구주와의관계=="가구주의 자녀","수도권\_가구주의 자녀",수도권$가구주와의관계))))경남<-subset(file,file$지역=="경남")경남$인구구분<-ifelse(경남$가구주와의관계=="가구주","경남\_가구주", ifelse(경남$가구주와의관계=="가구주의 배우자","경남\_가구주의 배우자", ifelse(경남$가구주와의관계=="가구주의 손주","경남\_가구주의 손주", ifelse(경남$가구주와의관계=="가구주의 자녀","경남\_가구주의 자녀",경남$가구주와의관계))))경북<-subset(file,file$지역=="경북")경북$인구구분<-ifelse(경북$가구주와의관계=="가구주","경북\_가구주", ifelse(경북$가구주와의관계=="가구주의 배우자","경북\_가구주의 배우자", ifelse(경북$가구주와의관계=="가구주의 손주","경북\_가구주의 손주", ifelse(경북$가구주와의관계=="가구주의 자녀","경북\_가구주의 자녀",경북$가구주와의관계))))충남<-subset(file,file$지역=="충남")충남$인구구분<-ifelse(충남$가구주와의관계=="가구주","충남\_가구주", ifelse(충남$가구주와의관계=="가구주의 배우자","충남\_가구주의 배우자", ifelse(충남$가구주와의관계=="가구주의 손주","충남\_가구주의 손주", ifelse(충남$가구주와의관계=="가구주의 자녀","충남\_가구주의 자녀",충남$가구주와의관계))))강원\_충북<-subset(file,file$지역=="강원/충북")충북$인구구분<-ifelse(강원\_충북$가구주와의관계=="가구주","강,축\_가구주", ifelse(강원\_충북$가구주와의관계=="가구주의 배우자","강원\_충북\_가구주의 배우자", ifelse(강원\_충북$가구주와의관계=="가구주의 손주","강원\_충북\_가구주의 손주", ifelse(강원\_충북$가구주와의관계=="가구주의 자녀","강원\_충북\_가구주의 자녀",강원\_충북$가구주와의관계))))전라\_제주<-subset(file,file$지역=="전라/제주")전라\_제주$인구구분<-ifelse(전라\_제주$가구주와의관계=="가구주","전,제\_가구주", ifelse(전라\_제주$가구주와의관계=="가구주의 배우자","전라\_제주\_가구주의 배우자", ifelse(전라\_제주$가구주와의관계=="가구주의 손주","전라\_제주\_가구주의 손주", ifelse(전라\_제주$가구주와의관계=="가구주의 자녀","전라\_제주\_가구주의 자녀",전라\_제주$가구주와의관계))))인구구분<-c(서울$인구구분,수도권$인구구분,경남$인구구분,경북$인구구분,충남$인구구분,충북$인구구분,전라\_제주$인구구분)table(인구구분)wordcloud2(data=table(인구구분),fontFamily = '나눔바른고딕',color = "random-light") |

1. **결과물**

**▶** 마지막으로 워드 클라우드를 이용해서 어떤 형태의 사람들이 가장 많은지 알아 보았다. 이 자료는 인구 수 자료와 비슷한 양상을 보였다. 가구주들 중에서는 전라/제주가 가장 많고 다음이 경남, 수도권 순이다. 가구주 이외에 나머지 데이터는 잘 보이지 않지만 하늘색으로 수도권\_가구주 자녀가 보인다. 즉 한부모 가정의 저소득 가구는 수도권이 가장 많은 것을 확인할 수 있다.  부록 (전체 코드)# (1) 분석 작업을 위해 자신의 작업 환경 및 데이터 준비# 및 한국복지패널 데이터 R-studio로 불러오기library(foreign)f<-read.spss(file="Koweps\_hpc10\_2015\_beta1.sav",to.data.frame=T)# (2) 한국복지패널 데이터(f)의 기초적인 탐색하기str(f) # 데이터 세트의 구성colnames(f) # 변수의 이름dim(f) # 행과 열의 개수head(f,3) # 일부 데이터만 보기 View(f) # 전체 데이터 보기summary(f) # 요약본 보기 # (3) 분석 작업에 필요한 패키지 또는 라이브러리 설치 및 로드library(dplyr)library(gdata)library(stringr)library(ggplot2)library(devtools)devtools::install\_github("lchiffon/wordcloud2")library(wordcloud2)# (4) 한국복지패널데이터에서 분석할 관심 문제를 제안하시오.# 문제: 저소득 가구와 지역간의 관계 # (5) 분석 작업할 한국복지패널데이터를 R-studio에서 작업할 수 있도록 # 변수를 정하고 데이터를 가공하시오 # 필요한 변수 정하기 default<-"h10"vector<-c("\_id","\_reg7","\_din","\_hc","01\_1","\_g2","\_g3","\_g4","01\_110")vector\_name<-c("가구번호","지역","가처분소득","가구구분","가구원수","가구주와의관계","성별","태어난연도","가구형태")# 필요한 변수만 가져와서 리스트 만들기file<-list()for(i in 1:9){ file<-c(file,f[paste(default,vector[i],sep='')])}# 리스트를 데이터 프레임으로 변환하고 열 이름 바꾸기 file<-as.data.frame(file)for(i in 1:9){ colnames(file)[i]<-vector\_name[i]}# 필요한 변수만 가져온 데이터 보기 View(file)# file 데이터 세트 가공하기 # 각 행의 데이터를 문자열로 보이게 하기file$"지역"<-factor(file$"지역",levels=c(1,2,3,4,5,6,7),labels=c("서울","수도권","경남","경북","충남","강원/충북","전라/제주"))file$"성별"<-factor(file$"성별",levels=c(1,2),labels=c("남","여"))file$"가구구분"<-factor(file$"가구구분",levels=c(1,2),labels=c("일반 가구","저소득층 가구"))file$"가구형태"<-factor(file$"가구형태",levels=c(1,2,3,4,5),labels=c("단독","모자/모녀","부자/부녀","조손가구","기타"))# "가구주와의관계"가 너무 복합함 table(file$"가구주와의관계")# "가구주와의관계" 데이터를 (함축해서) 문자열로 보이게 하기file$"가구주와의관계"<-ifelse(11<=file$"가구주와의관계" & file$"가구주와의관계"<=16,"가구주의 자녀", ifelse(111<=file$"가구주와의관계" & file$"가구주와의관계"<=166,"가구주의 손주", ifelse(21<=file$"가구주와의관계" & file$"가구주와의관계"<=26,"가구주 자녀의 배우자", ifelse(file$"가구주와의관계"==10,"가구주", ifelse(file$"가구주와의관계"==20,"가구주의 배우자",file$"가구주와의관계")))))file$"가구주와의관계"<-as.factor(file$"가구주와의관계")# 가구형태가 "기타"인 데이터가 너무 복잡하므로 삭제 # (부모,자녀가 같이 사는 가구, 룸메이트끼리 사는 가구, 친인척이랑 사는 가구 등) 삭제 table(file$"가구형태")file<-filter(file,file$"가구형태" != "기타")# 가구구분이 "저소득가구" 데이터만 남겨놓음table(file$"가구구분")file<-filter(file,file$"가구구분" != "일반 가구")# "저소득 가구"데이터 밖에 없으니 "가구구분" 변수 삭제file<-select(file,-가구구분)# 데이터 정제 (이상치 처리)file<-filter(file,file$가처분소득>0) # 소득이 마이너스인 가구 제거 file$"가구주와의관계"<-drop.levels(file$"가구주와의관계") # 값이 0인 레벨 제거file$"가구형태"<-drop.levels(file$"가구형태") # 값이 0인 레벨 제거file$"가구구분"<-drop.levels(file$"가구구분") # 값이 0인 레벨 제거# (6) 가공된 한국복지패널데이터(file) 의 기초적인 탐색하기str(file) # 데이터 세트의 구성 colnames(file) # 변수의 이름dim(file) # 행과 열의 개수head(file,3) # 일부 데이터만 보기 View(file) # 전체 데이터 보기summary(file) # 요약본 보기 # (7) 데이터 분석: 탐색, 시각화, 분석# 데이터를 편하게 사용하기 위해 attach 함수를 사용attach(file)# 지역 별 저소득 가구의 인구 수people\_count<-table(지역)pie(people\_count,xlab="지역 별 저소득 가구의 인구 수")# 지역 별 저소득 가구의 인구 수 (성별 포함)people\_count2<-table(성별,지역)barplot(people\_count2,xlab="지역 별 저소득 가구의 인구 수",col=c("red","darkblue"),beside=T)legend("topleft",levels(성별),lty=c(1,1),col=c("red","blue"))# 지역 별 저소득 가구원 수people\_fcount<-table(가구원수,지역)barplot(people\_fcount,xlab="지역 별 저소득 가구원 수",col=c("red","orange","yellow","green","blue"),beside=T)legend("topleft",levels(as.factor(가구원수)),lty=c(1,1,1,1,1),col=c("red","orange","yellow","green","blue"))# 지역 별 저소득 가구의 가구형태 people\_type<-table(가구형태,지역)barplot(people\_type,xlab="지역 별 저소득 가구의 가구형태",col=c("red","orange","yellow","green"),beside=F)legend("topleft",levels(가구형태),lty=c(1,1,1,1),col=c("red","orange","yellow","green"))# 지역 별 저소득 가구의 소득 people\_money<-as.data.table(summarise(group\_by(file, 지역), mean(가처분소득)))colnames(people\_money)<-c("지역","가처분소득\_평균")ggplot(people\_money, aes(지역,가처분소득\_평균))+geom\_bar(stat="identity",fill="green")# 지역 별 저소득 가구의 나이대 file$나이대<-ifelse(태어난연도<=1950,"노년층", ifelse(태어난연도<=1970,"중년층", ifelse(태어난연도<=1990,"장년층", ifelse(태어난연도<=2005,"청년층","아동층"))))people\_age<-table(file$나이대,지역)barplot(people\_age,xlab="지역 별 저소득 가구의 나이대",col=c("red","orange","yellow","green","blue"),beside=T,horiz=T)legend("topright",levels(as.factor(file$나이대)),lty=c(1,1,1,1,1),col=c("red","orange","yellow","green","blue"))# 지역 별 저소득 가구의 인구 구분 서울<-subset(file,file$지역=="서울")서울$인구구분<-ifelse(서울$가구주와의관계=="가구주","서울\_가구주", ifelse(서울$가구주와의관계=="가구주의 배우자","서울\_가구주의 배우자", ifelse(서울$가구주와의관계=="가구주의 손주","서울\_가구주의 손주", ifelse(서울$가구주와의관계=="가구주의 자녀","서울\_가구주의 자녀",서울$가구주와의관계))))수도권<-subset(file,file$지역=="수도권")수도권$인구구분<-ifelse(수도권$가구주와의관계=="가구주","수도권\_가구주", ifelse(수도권$가구주와의관계=="가구주의 배우자","수도권\_가구주의 배우자", ifelse(수도권$가구주와의관계=="가구주의 손주","수도권\_가구주의 손주", ifelse(수도권$가구주와의관계=="가구주의 자녀","수도권\_가구주의 자녀",수도권$가구주와의관계))))경남<-subset(file,file$지역=="경남")경남$인구구분<-ifelse(경남$가구주와의관계=="가구주","경남\_가구주", ifelse(경남$가구주와의관계=="가구주의 배우자","경남\_가구주의 배우자", ifelse(경남$가구주와의관계=="가구주의 손주","경남\_가구주의 손주", ifelse(경남$가구주와의관계=="가구주의 자녀","경남\_가구주의 자녀",경남$가구주와의관계))))경북<-subset(file,file$지역=="경북")경북$인구구분<-ifelse(경북$가구주와의관계=="가구주","경북\_가구주", ifelse(경북$가구주와의관계=="가구주의 배우자","경북\_가구주의 배우자", ifelse(경북$가구주와의관계=="가구주의 손주","경북\_가구주의 손주", ifelse(경북$가구주와의관계=="가구주의 자녀","경북\_가구주의 자녀",경북$가구주와의관계))))충남<-subset(file,file$지역=="충남")충남$인구구분<-ifelse(충남$가구주와의관계=="가구주","충남\_가구주", ifelse(충남$가구주와의관계=="가구주의 배우자","충남\_가구주의 배우자", ifelse(충남$가구주와의관계=="가구주의 손주","충남\_가구주의 손주", ifelse(충남$가구주와의관계=="가구주의 자녀","충남\_가구주의 자녀",충남$가구주와의관계))))강원\_충북<-subset(file,file$지역=="강원/충북")강원\_충북$인구구분<-ifelse(강원\_충북$가구주와의관계=="가구주","강,축\_가구주", ifelse(강원\_충북$가구주와의관계=="가구주의 배우자","강원\_충북\_가구주의 배우자", ifelse(강원\_충북$가구주와의관계=="가구주의 손주","강원\_충북\_가구주의 손주", ifelse(강원\_충북$가구주와의관계=="가구주의 자녀","강원\_충북\_가구주의 자녀",강원\_충북$가구주와의관계))))전라\_제주<-subset(file,file$지역=="전라/제주")전라\_제주$인구구분<-ifelse(전라\_제주$가구주와의관계=="가구주","전,제\_가구주", ifelse(전라\_제주$가구주와의관계=="가구주의 배우자","전라\_제주\_가구주의 배우자", ifelse(전라\_제주$가구주와의관계=="가구주의 손주","전라\_제주\_가구주의 손주", ifelse(전라\_제주$가구주와의관계=="가구주의 자녀","전라\_제주\_가구주의 자녀",전라\_제주$가구주와의관계))))인구구분<-c(서울$인구구분,수도권$인구구분,경남$인구구분,경북$인구구분,충남$인구구분,강원\_충북$인구구분,전라\_제주$인구구분)wordcloud2(data=table(인구구분),fontFamily = '나눔바른고딕',color = "random-light") |  |
|  |  |