빅데이터 기말고사

20대 업종형태 트렌드와 지역별 차이 분석

컴퓨터공학과

201533689 고수균

20대 업종형태 트렌드와 지역별 차이 분석

Sukun

2018-06-18

1. 데이터 정의 및 대상 분석 요약

시대가 흘러감에 따라 노동시장에서는 연령별 분리현상이 일어나고 있습니다 1). 현 젊은 세대 노동시장 형태의 지표라고 할 수 있는 20대의 업종 종사 형태와 지역별 차이를 통하여 최근 노동시장의 상황을 분석해 보고자 하였습니다. 본 분석은 한국보건사회연구원에서 가구의 경제활동을 연구해 발간한 2015년도 한국복지패널데이터 중 통합 데이터인 Koweps_hpc10_2015_beta1.sav를 이용하였습니다. 분석에 활용한 변수는 총 3가지로, 아래와 같습니다. (1) h10_g4 : 출생년도. 응답자의 태어난 연도에 해당하는 변수로서 int형으로 연도에 해당하는 숫자가 입력되어 있습니다. 20대를 파악할 때 활용하였습니다. (2) h10_eco8 : 업종코드. 21개의 대분류, 이를 01부터 99까지 중 몇 개의 번호가 누락된 76개의 중분류로 나누어 업종을 나타내고 있습니다. 업종의 형태를 파악할 때 활용하였습니다. (3) h10_reg7 : 지역코드. 지역별 데이터로서, 7가지로 구분하였고 1부터 7까지의 코드를 통해 그를 나타내고 있습니다. 지역별로 분리할 때 활용하였습니다.

2. R스크립트

// package install 항목들은 생략하였습니다. 과정과 크게 관련 없는 실행문들은 생략하였습니다. library(foreign) library(dplyr) library(ggplot2) library(readxl)

raw_welfare <- read.spss(file = “Koweps_hpc10_2015_beta1.sav”, to.data.frame = T) welfare <- raw_welfare

str(welfare) welfare <- rename(welfare, birth = h10_g4, business_category = h10_eco8, code_region = h10_reg7) // 활용할 변수들을 쓰임새에 맞는 변수명으로 재선언 하였습니다.

summary(welfare$birth)table(is.na(welfare$birth)) welfare$birth < \quad ifelse(((2018 \quad welfare$birth)<15 | (2018 - welfare$birth) > 60), NA, welfare$birth) // 데이터 1차 정제 - 근로기준법 의거, 2015년 합법적인 근로 연령은 만 15세부터 만 60세입니다 2), 3). 이에 1차적으로 이외 데이터들을 결측치로 처리하여 근로 가능 인구 데이터를 정제하였습니다.

welfare$age < \quad 2018 \quad welfare$birth +1 summary(welfare$age)

welfare <- welfare %>% mutate(age_group = ifelse(age < 20, “10대”, ifelse(age < 30 , “20대”, ifelse(age < 40, “30대”, ifelse(age < 50, “40대”, “50대~정년퇴임”))))) // 데이터 2차 정제를 위한 전처리 - 20대에 해당하는 데이터를 얻기 위하여, 나이별로 연령대를 나누어 연령대 변수를 추가하였습니다.

welfare$category_name \quad < \quad ifelse(welfare$business_category <=3, “농림업”, ifelse(welfare$business_category \quad <= 8, , ifelse(welfare$business_category <= 33, “제조업”, ifelse(welfare$business_category <= 36, , , , ifelse(welfar
e$business_category
<= 39, “환경복원업”, ifelse(welfare$business_category \quad <= \quad 42, , ifelse(welfare$business_category <= 47, “도매 및 소매업”, ifelse(welfare$business_category \quad <= \quad 52, , ifelse(welfare$business_category <= 56, “숙박, 음식점 및 주점업”, ifelse(welfare$business_category \quad <= \quad 63, , , , ifelse(welfare$business_category <= 66, “금융 및 보험업”, ifelse(welfare$business_category \quad <= \quad 69, , ifelse(welfare$business_category <= 73, “전문, 과학 및 기술 서비스업”, ifelse(welfare$business_category \quad <= \quad 75, , ifelse(welfare$business_category <= 84, “공공행정, 국방 및 사회보장 행정업”, ifelse(welfare$business_category <= 85, , ifelse(welfare$business_category <= 87, “보건 및 사회복지 사업”, ifelse(welfare$business_category <= 91, , , ifelse(welfare$business_category <= 96, “협회 및 단체, 수리 및 기타 개인 서비스업”, ifelse(welfare$business_category <= 98, “기타 자가소비 생산활동”, “국제 및 외국기관”, NA))))))))))))))))))))) // 업종 코드 대분류에 따라 업종을 묶어주었습니다. 텍스트 마이닝에 활용하고자 명사가 잘 드러나 있는 분류명이 입력될 수 있도록 하였습니다.

library(KoNLP) library(dplyr) useNIADic()

twen <- welfare %>% filter(age_group==“20대”) %>% select(category_name) table(twen$category_name) write.csv(twen, file=“twen.csv”) // 20대에 해당하는 사람들의 업종명 변수에 입력된 대분류 업종명들을 csv파일로 저장하였습니다. 20대에

해당하는 각 사람들의 종사 업종에 대한 대분류 항목이 텍스트로 저장되어 이를 분석하면, 텍스트에 등장하는 명사의 빈도에 따라 유의한 분석을 할 수 있을 것이라 생각하였습니다. twentxt <- readLines(“twen.csv”) // 분석을 위해 csv파일의 텍스트를 읽어옵니다.

library(stringr)

twentxt <- str_replace_all(twentxt, “ \ W”, ” “) // 텍스트 파일의 특수문자들을 모두 제거해 주었습니다.

twennouns <- sapply(twentxt, extractNoun, USE.NAMES = F) // 텍스트 파일의 한글만 뽑아낼 수 있도록 해주었습니다 4). csv파일 형태로 저장될 때 불필요한 번호들이 “숫자” 형태로 저장되어 문자로 저장되므로 이를 처리하였습니다. jobcount <- table(unlist(twennouns)) twen_job <- as.data.frame(jobcount, stringAsFactors = F) twen_job <- rename(twen_job, word = Var1, freq = Freq)

twen_job$word < \quad as.character(twen_j ob$word) twen_job <- filter(twen_job, nchar(word) >= 2) // 업종 대분류명을 보면, 한 글자에 해당하는 단어는 의미가 없음을 파악할 수 있습니다. 이에 두 글자 이상의 단어만 처리할 수 있도록 하였습니다.

summary(twen_job$freq)//.boxplot(twen_j ob$freq)$stats // 요약 통계량과 boxplot을 이용해 파악한 결과 “숫자”에 해당하는 데이터들이 아직 남아있는 것을 파악하였습니다.

twen_job <- twen_job %>% filter(freq > 1) // 불필요한 데이터들은 대부분 1의 빈도로 인식되어 있었고, 이는 경향 분석에 지장을 주지 않는다고 판단하여 1보다 빈도가 높은 데이터만 추출하여 업종 형태에 대한 분석을 진행하였습니다.

// 추출 결과 명사 데이터가 많아 차트나 그래프 형태로는 시각화가 힘들다고 판단, 워드클라우드를 이용하여 시각화하였습니다. library(wordcloud) library(RColorBrewer)

pal <- brewer.pal(8, “Dark2”)

set.seed(1234) wordcloud(words = twen_job$word, freq = \quad twen_j ob$freq, min.freq = 3, max.words = 71, random.order = F, rot.per = .1, scale = c(4, 0.3), colors = pal)

twen_region <- welfare %>% filter(age_group==“20대”) %>%
select(business_category, code_region) // 지역별로 차이를 파악하기 위해 20대 데이터의 업종 코드와 지역코드 변수를 추출하였습니다. 앞서 분석 과정을 통해 업종 형태 경향에 대한 분석은 끝났으므로 특정 경향에 대한 업종 코드를 기준으로 지역별 차이를 분석하고자 하였습니다.

table(is.na(twen_region)) twen_region <- twen_region %>% filter(!is.na(business_category)) table(is.na(twen_region)) // 업종 코드가 비어있는 경우 분석에 도움이 되지 않으므로 결측치 처리를 하였습니다.

seoul_business <- twen_region %>% filter(code_region==1) %>% select(business_category) // 각 지역에 해당하는 데이터를 추출하여 따로 저장하고, 파생된 데이터는 특정 지역에 해당하는 사람들의 업종명만 입력되어 있으므로 업종 변수만 추출하였습니다.

seoul_service_job <- seoul_business %>% filter(business_category %in% c(58, 59, 60, 61, 62, 63, 70, 71, 72, 73, 74, 75, 85, 90, 91, 94, 95, 96)) %>% summarise(“인원”=n()) // 20대 업종 경향에 대한 분석 결과 가장 많은 형태인 “서비스”가 들어간 대분류의 하위 업종 코드에 해당하는 업종 코드들의 빈도를 계산하였고, 이를 통해 20대 응답자 중 “서비스” 형태의 업종에 종사하는 사람들의 수를 파악하였습니다. 이하 나머지 6개의 지역에 대하여도 위와 같은 방식으로 진행하였습니다.

seoul_service_job <- seoul_service_job %>% mutate(“지역” = “서울”) // 시각화를 위해 지역이라는 파생변수를 추가하였습니다. 나머지 6개 지역도 동일한 방식으로 진행하였습니다.

service_job <- bind_rows(seoul_service_job, gyeonggi_service_job, gyeongsangnam_service_job, gyeongsangbuk_service_job, chungcheongnam_service_job, chungcheongbuk_gangwon_service_job, jeolla_jeju_service_job) // 각 지역에 따른 인원 데이터들을 합쳐서 시각화 할 수 있도록 하였습니다.

ggplot(data = service_job, aes(x = reorder(지역, 인원), y = 인원)) + geom_col() + coord_flip() // 지역에 따른 인원 데이터를 인원 순으로 시각화하여 나타내었습니다.

library(kormaps2014)

korpop1 <- korpop1

name <- c(“서울”, “경남”, “경북”, “경기”, “전라/제주”, “충남”, “경남”, “충남”, “경기”, “충북/강원”, “충북/강원”, “충남”, “전라/제주”, “전라/제주”, “경북”, “경남”, “전라/제주”) // 지도 시각화를 통해 지역별 서비스업종 종사 인구 분포가 확연히

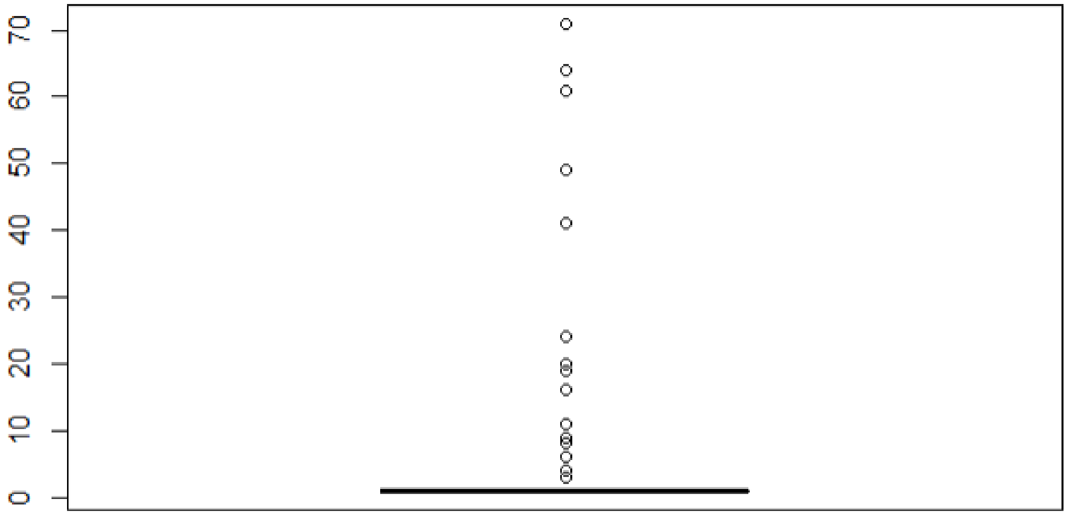
드러날 수 있도록 하였습니다. 이 과정에서 korpop1 데이터는 시/도별로 되어 있는 반면, 본 분석에 사용한 데이터는 7개의 지역으로 분류되어 있습니다. 이에 각 지역에 해당하는 시/도 코드는 모두 각 지역 데이터를 넣어서 지도에 매핑될 때 같은 밀도로 표현될 수 있도록 하였습니다. pop <- c(19, 10, 3, 22, 12, 8, 10, 8, 22, 5, 5, 8, 12, 12, 3, 10, 12) code <- c("11", "21", "22", "23", "24", "25", "26", "29", "31", "32", "33", "34", "35", "36", "37", "38", "39")

korpop <- data.frame(name, pop, code)

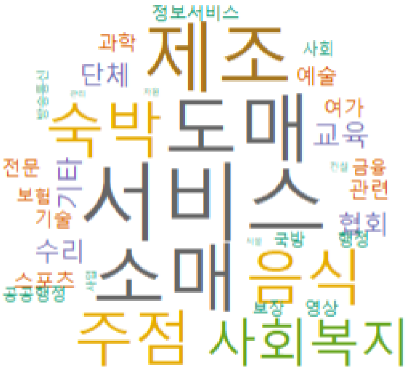
ggplot(korpop,aes(fill = pop, map_id=code, tooltip = name), interactive = T) + #만든 korpop을 지도에 표시한다. geom_map(map=kormap1,colour="black",size=0.1) + expand_limits(x=kormap1$long, y = kormap1$lat) + scale_fill_gradientn(colours=c('white'
+ ggtitle("2015년도 지역별 서비스 형태 업종 종사 분포") + coord_map()

3. 과정과 결과

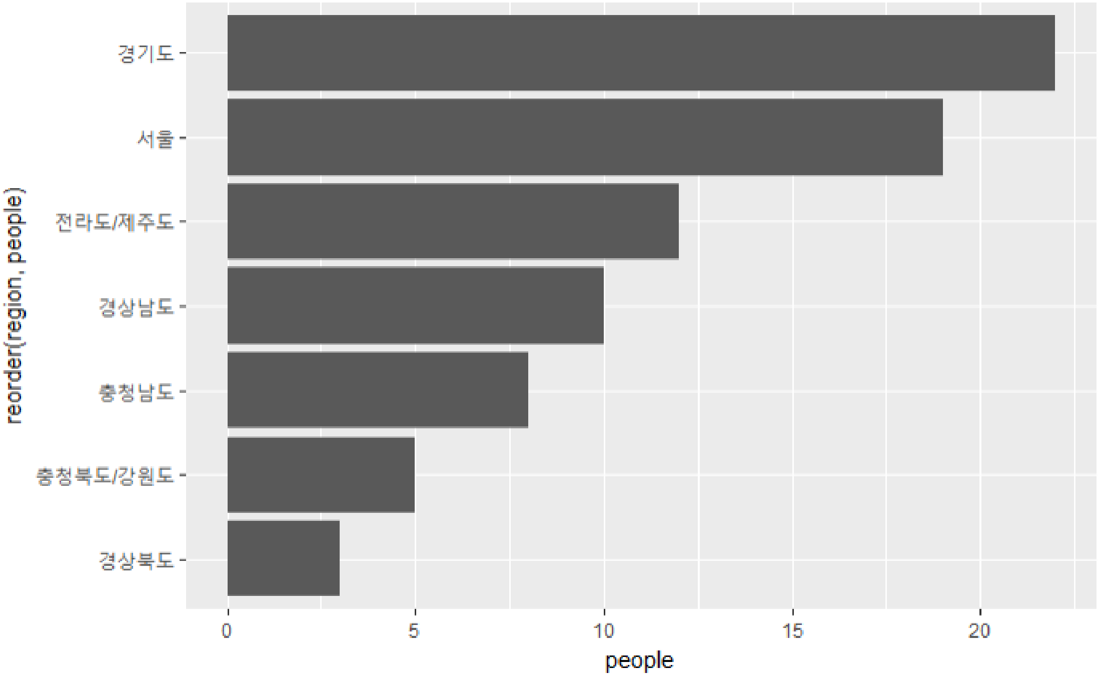
본 분석은 현 젊은 세대 노동시장 형태의 지표라고 할 수 있는 20대의 업종 종사 형태와 지역별 차이를 통하여 최근 노동시장의 상황을 분석해 보고자 하였던 분석입니다. 20대가 종사하는 업종들에게서 나타나는 공통적인 형태를 분석하고자 분류 업종명 내 명사 빈도을 이용한 텍스트 마이닝을 통하여 분석을 진행하였습니다.



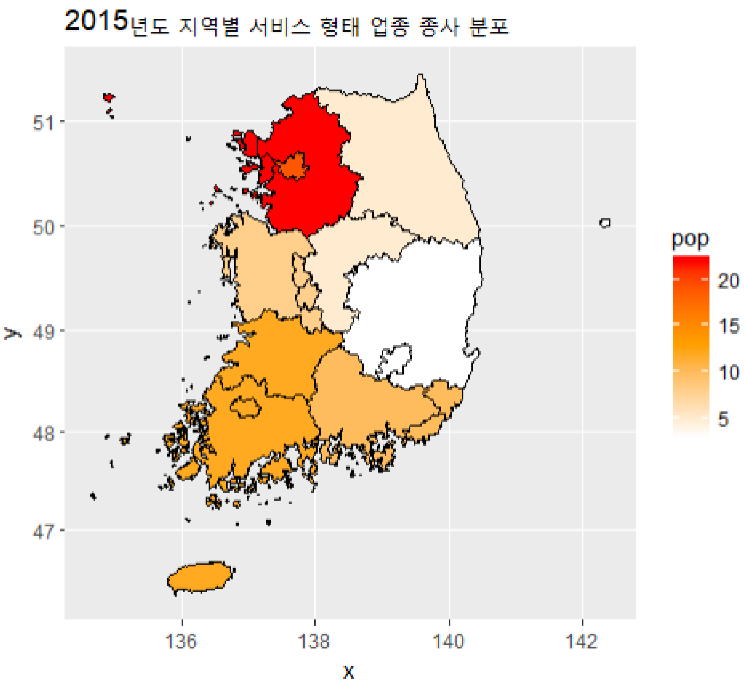
텍스트 마이닝을 통해 도출된 통계치에서, 처리를 통해 완전히 걸러지지 않은 불필요한 값들이 남아있음을 파악하게 되었고, 보다 간경한 경향성 파악을 위하여 불필요한 통계치의 값들을 빈도 임계값을 이용하여 제거하였습니다. 데이터를 지나치게 손상시키지 않는 한에서 빈도 임계값을 설정하였고, 1로 설정하였습니다.



텍스트 마이닝을 통해 최종적으로 도출된 결과를 파악하기 쉽도록 워드클라우드를 이용하여 시각화하였고, 20대 업종 형태의 경향성을 파악할 수 있었습니다.



업종 형태의 경향성 파악 결과를 바탕으로 지역별로 그러한 트렌드에 얼마나 반응하고 있는지/ 차이가 있는지 파악하기 위하여 지역별로 인원을 파악하였고, 시각화하였습니다.



마지막으로 지역별 인원을 지도에 매핑하여 20대 전체 업종 형태 트렌드에 대한 지역별 차이/민감도를 한눈에 알기 쉽도록 하였습니다.

4. 해석과 결론

20대 업종 종사 형태에 대한 분석 결과 현 20대는 서비스 형태의 업종에 주로 종사한다는 것을 파악할 수 있었습니다. 업종의 형태가 주로 서비스업에 집중되는 방향으로 노동시장의 구조가 형성될 것으로 전망할 수 있었고, 기존 제조나 도/소매 업종으로부터 이 새로운 업종 형태 트렌드가 현재 우리나라의 유행 또한 나타내주고 있다고 생각되었습니다. 또한, 지역별로 이러한 트렌드에 대한 민감도 차이를 살펴보았을 때, 수도권이 가장 민감하며, 강원도와 경상북도 지역에서는 이러한 흐름에 대한 민감도가 떨어진다는 것을 파악할 수 있었습니다. 강원도의 경우, 강원도의 낮은 인구 밀도가 기존 주 업종에 대한 유입이나 진출을 방해하는 요인으로 작용하며 5), 경상남도의 경우 울산의 공장지대 등 기존의 제조업과 같은 이미 형성된 업종층이 쉽게 감소되지 않기 때문이라고 생각합니다 6).

5. 참고문헌

1) 박경숙,「노동시장의 고령화와 업종/직종에서의 연령분리현상」,『노동정책연구』제2권 2호, 한국노동연구원, 2002, 81~82면.

2) 노동정보웹사이트, “알바를 할 수 있는 나이와 준비할 사항은?”, 2014.06.08, http://www.nodong.or.kr/alba_sense/1462641, 2018.06.18.

3) 이형,「정년연장에 따른 노동법상의 쟁점과 과제」,『산업관계연구』24권 1호, 한국고용노사관계학회, 2014, 21~22면.

4) 데이터과학자를 꿈꾸며 티스토리 블로그, [R] R 9일차, 2018.05.23, http://datamod.tistory.com/95, 2018,06,18.

5) 한국은행 강원본부,「강원경제의 현황과 발전방향」,한국은행, 2008, 3~4면.

6) 박병주,「경상남도 동북아 핵심 물류거점 구축을 위한 장기전략 수립용역」, 온-나라 정책연구, 2014, 17~21면.