

# Election Visualization

Jaeseok Jeong

Department of Statistics, SKKU

August 29, 2018

## 1. 선거데이터 전처리

```
# 단위 : 시군구별 -> 시군구별 합계만 추출

election_data <- election_data %>%
  dplyr::filter(읍면동명 == "합계")

# 구시군명 중 띄어쓰기 있는 부분 붙여쓰기로 변경
# 나중에 shp.file과 결합위해 변경
# example : "청주시 상당구" -> "청주시상당구"

election_data$구시군명 <- election_data$구시군명 %>%
  stringr::str_replace(pattern = ' ', replacement = '')

# ,로 분리된 숫자 -> , 지우고 chracter로 설정된 속성을 numeric으로 변경

for(i in 5:ncol(election_data)) {

  election_data[,i] <- as.numeric(election_data[,i] %>%
    stringr::str_replace(pattern = ',', replacement = ''))
}
```

## 2. Top 5 후보만 뽑음

문재인, 홍준표, 안철수, 유승민, 심상정

5명 이외의 후보의 시군구별 지지율은 큰 의미 없음

```
election_top_5 <- election_data %>%
  dplyr::select(1:11)

colnames(election_top_5)

## [1] "시도명"           "구시군명"         "읍면동명"
## [4] "투표구명"         "선거인수"         "투표수"
## [7] "더불어민주당\n문재인" "자유한국당\n홍준표" "국민의당\n안철수"
## [10] "바른정당\n유승민"  "정의당\n심상정"
```

```

# 변수이름 영어로 변경

colnames(election_top_5)[6:11] <- c("vote_sum", "moon", "hong", "ahn", "yoo", "sim")

#각 후보들의 지역별 지지율(prop)

election_top_5 <- election_top_5 %>%
  dplyr::mutate(prop_moon = moon/vote_sum,
                prop_hong = hong/vote_sum,
                prop_ahn = ahn/vote_sum,
                prop_yoo = yoo/vote_sum,
                prop_sim = sim/vote_sum)

```

### 3. 지역데이터(shp.file) 전처리

```

region_data <- rgdal::readOGR(dsn = 'C:/Users/82104/Desktop/map',
                              layer = 'TL_SCCO_SIG',
                              encoding = 'UTF-8')

```

```

## OGR data source with driver: ESRI Shapefile
## Source: "C:\Users\82104\Desktop\map", layer: "TL_SCCO_SIG"
## with 250 features
## It has 3 fields

```

```

region_data@data$id <- rownames(region_data@data)

region_dataframe <- ggplot2::fortify(region_data)

```

```

## Regions defined for each Polygons

```

```

final_region_data <- merge(x = region_dataframe,
                          y = region_data@data,
                          by = 'id',
                          all.x = TRUE)

```

```

# 구시군명 중 띄어쓰기 있는 부분 붙여쓰기로 변경
# example : "청주시 상당구" -> "청주시상당구"

```

```

final_region_data$SIG_KOR_NM <- final_region_data$SIG_KOR_NM %>%
  stringr::str_replace(pattern = ' ', replacement = '')

```

```

# 데이터 정렬

```

```

final_region_data <- final_region_data %>%
  dplyr::arrange(id, order)

```

```

# 데이터 속성 확인

```

```

dim(final_region_data)

```

```

## [1] 1337823      10

```

```
head(final_region_data,3)
```

```
##   id      long      lat order hole piece group SIG_CD SIG_ENG_NM SIG_KOR_NM
## 1  0 956615.5 1953567     1 FALSE     1   0.1  11110 Jongno-gu   종로구
## 2  0 956621.6 1953565     2 FALSE     1   0.1  11110 Jongno-gu   종로구
## 3  0 956626.2 1953564     3 FALSE     1   0.1  11110 Jongno-gu   종로구
```

## 4. 지역데이터 규모 축소

행이 1,337,823개 - 전체데이터 사용하는 경우 실행시간 오래 걸림, 부속지역(섬) 때문에 시각화 더러움

```
# 행길이 1/10으로 축소
```

```
subset_region_data <- final_region_data[final_region_data$order %% 10 == 1, ]
```

```
# 부속지역을 제외
```

```
subset_region_data <- subset_region_data %>%
  dplyr::filter(piece == 1)
```

## 5. 지역데이터 & 선거데이터 결합

```
election_result <- merge(x = subset_region_data,
  y = election_top_5,
  by.x = 'SIG_KOR_NM',
  by.y = '구시군명',
  all.x = TRUE)
```

```
# 데이터 정렬
```

```
election_result <- election_result %>%
  dplyr::arrange(id, order)
```

## 6. 결과 시각화

### 문재인 후보 시군구별 지지율

```
ggplot(data = election_result, mapping = aes(x = long, y = lat, group = group)) +  
  geom_polygon(mapping = aes(fill = prop_moon), color = 'black') +  
  coord_fixed() +  
  scale_x_continuous(breaks=seq(700000, 1400000, 200000)) +  
  scale_fill_gradient(low = 'yellow', high = 'red')
```

