

# finance

lee

2021년 8월 31일

금융데이터 경진대회 코드 제출

## 0. Preparation

- 0.1 패키지 불러들이기
- 0.2 서울시 지역단위 '소득', '지출', '금융자산' 정보 데이터 불러들이기

```
fin_2103_o <-
read_xlsx('C:/Users/Lee/Downloads/finance_data202103.xlsx')
fin_2103 <- fin_2103_o[,-c(1,3,5)]  
  
str(fin_2103)
고령자_o <- read.delim('C:/Users/Lee/Downloads/고령자_2.txt', header =
F, skip = 5, fileEncoding = "UTF-8")
colnames(고령자_o) [c(2,3,7)] <- c('자치구', '행정동', '고령자계')
고령자 <- 고령자_o %>% select(행정동, 고령자계) %>% filter(행정동 != '소계')  
  
장애인_o <- read.delim('C:/Users/Lee/Downloads/장애인_2.txt', header =
F, skip = 5, fileEncoding = "UTF-8")
colnames(장애인_o) [2:4] <- c('자치구', '행정동', '장애인계')
장애인 <- 장애인_o %>% select(행정동, 장애인계) %>% filter(행정동 != '소계')  
  
기초생활_o <- read.delim('C:/Users/Lee/Downloads/기초생활.txt', header =
F, skip = 4, fileEncoding = "UTF-8")
colnames(기초생활_o) [2:4] <- c('자치구', '행정동', '기초생활계')
기초생활<- 기초생활_o %>% select(행정동, 기초생활계) %>% filter(행정동 !=
's소계')  
  
주민등록인구_o <- read.delim('C:/Users/Lee/Downloads/주민등록인구.txt',
header = F, skip = 3, fileEncoding = "UTF-8")
colnames(주민등록인구_o) [2:5] <- c('자치구', '행정동', '세대', '주민등록계')
주민등록인구<- 주민등록인구_o %>% select(행정동, 주민등록계) %>%
filter(행정동 != '소계')  
  
colSums(고령자 == 0) / nrow(고령자)
colSums(고령자 == 0) / nrow(고령자)
• 0.3 결측치 확인 및 0값 확인
sum(is.na(fin_2103))
```

```

## [1] 0
zero_ratio <- as.data.frame(colSums(fin_2103 == 0)) [1] / 211886
zero_ratio
##                                     colSums(fin_2103 == 0)
## 지역구                               0.0000000
## 법정동                               0.0000000
## 나이                                 0.0000000
## 성별                                 0.0000000
## 직장인여부                           0.9609885
## 급여입금                             0.9609885
## 가맹점매출입금                       0.9994856
## 연금입금                             0.9612669
## 총소비금액                           0.2412807
## 총수신금액                           0.0000000
## 예적금금액                           0.6448656
## 신탁금액                             0.9812635
## 수익증권금액                         0.9891970
## 신용대출금액                         0.9887675
## 담보대출금액                         0.9793521
## 주택대출금액                         0.9990514
## 전세자금대출금액                     0.9961489

• 0.4 data type 정리
fin_2103$지역구 <- as.factor(fin_2103$지역구)
fin_2103$법정동 <- as.factor(fin_2103$법정동)
fin_2103$직장인여부 <- as.factor(fin_2103$직장인여부)

• 0.5 데이터 정제 및 파생변수 생성
# 총소비금액이 0인 데이터 제외
fin_2103 <- fin_2103 %>% filter(총소비금액 != 0)

# 기준변수로 총소득, 총소비, 투자자산, 대출을 의미하는 파생변수 생ㅅ
fin_2103$총소득 <- fin_2103$급여입금 + fin_2103$가맹점매출입금 +
fin_2103$연금입금 + fin_2103$총수신금액
fin_2103$총소비 <- fin_2103$총소비금액
fin_2103$투자자산 <- fin_2103$예적금금액 + fin_2103$신탁금액 +
fin_2103$수익증권금액
fin_2103$대출 <- fin_2103$신용대출금액 + fin_2103$담보대출금액 +
fin_2103$주택대출금액 + fin_2103$전세자금대출금액

• 0.6 법정동별 각 칼럼의 평균 산출
fin_2103_sml <- fin_2103[,c('법정동', '총소득', '총소비', '투자자산', '대출')]
str(fin_2103_sml)

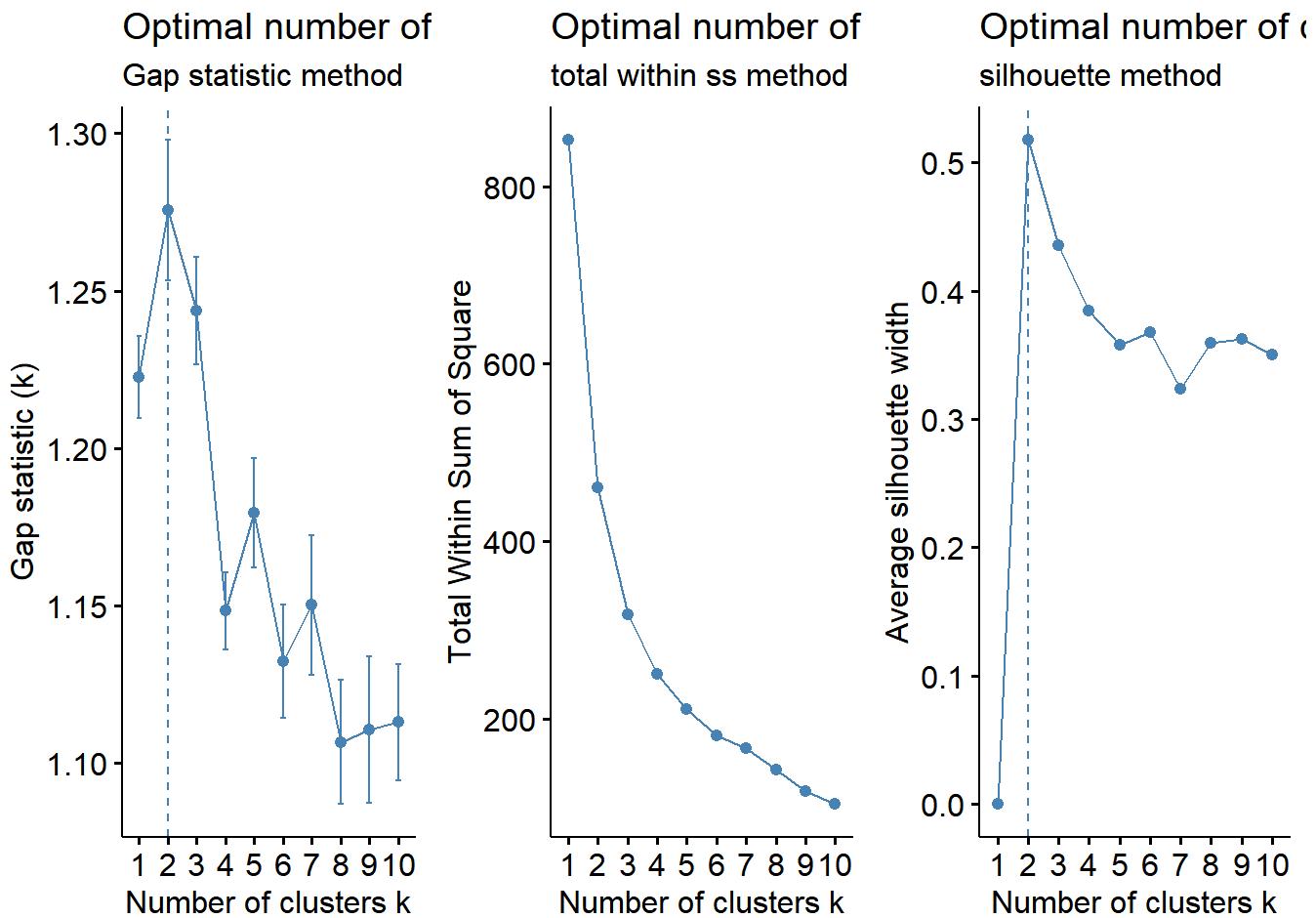
#동별 인당 평균
fin_2103_c <- fin_2103_sml %>% group_by(법정동) %>%
  summarise(count = n(), 총소득 = sum(총소득)/count, 총소비 =
sum(총소비)/count,
            투자자산 = sum(투자자산)/count,
            대출 = sum(대출)/count) %>% select(-count)

```

---

## 1. k-means clustering

- 1.1 k값 선정

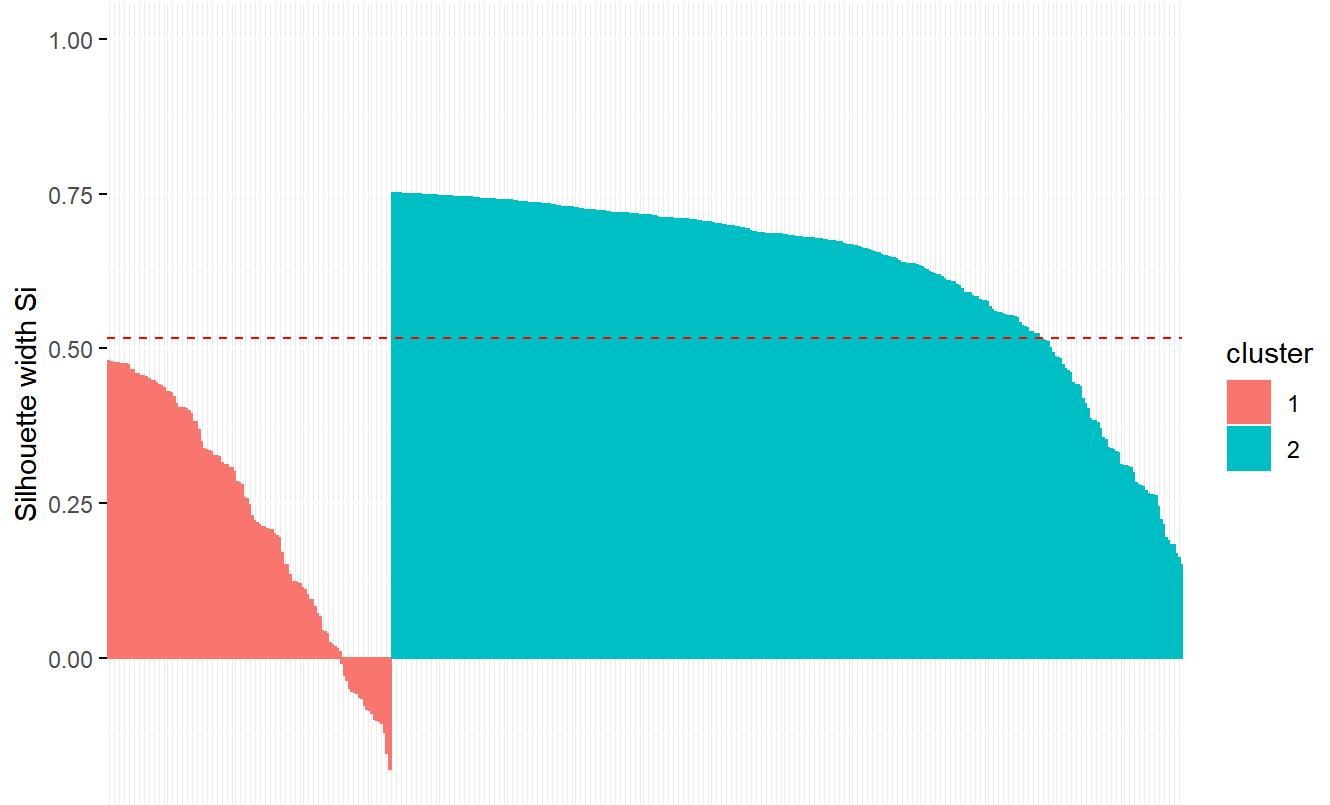


차원의 저주를 방지하고자 최소한의 칼럼으로 클러스터링을 진행하고자 함. gap statistic, total within sum of square, silhouette method로 최적의 k값을 선정함.

- 1.2 군집분석

```
kmn <- kmeans(scale(fin_2103_c[, c(2,4)]), centers = 2)
sil <- silhouette(kmn$cluster, dist(scale(fin_2103_c[, c(2,4)])))
fviz_silhouette(sil)
```

Clusters silhouette plot  
Average silhouette width: 0.52



- 1.2 군집별 특성 탐색

```

rst_o_2 <- cbind(fin_2103_c, kmn$cluster) %>% as_tibble()
rst_o_2
## # A tibble: 427 x 6
##   법정동     총소득    총소비    투자자산    대출 `kmn$cluster` 
##   <fct>      <dbl>     <dbl>     <dbl>     <dbl>       <int>
## 1 가락동  5960274. 1708744.  6153287.  7821802.      2
## 2 가리봉동 2189815. 1245759.  1158798.  850861.      2
## 3 가산동  4429734. 1550864.  3604310.  17662936.     2
## 4 가양동  5672956. 1611462.  6278572.  9035706.     2
## 5 가회동  5713882. 1908611.  5966100.  2993162.     2
## 6 갈월동  3273659. 1554874.  1263520.     0         2
## 7 갈현동  3839204. 1505525.  4292557.  676314.      2
## 8 강일동  4350206. 1581729.  2431666.  8009479.     2
## 9 개봉동  3944097. 1445056.  2855666.  4761710.     2
## 10 개포동  7296313. 1857157. 10983874. 11621110.     1
## # ... with 417 more rows
colnames(rst_o_2)[6] <- 'cluster'
rst_o_2$cluster <- as.factor(rst_o_2$cluster)

rst_o_2 %>% group_by(cluster) %>% summarise(평균총소득 =

```

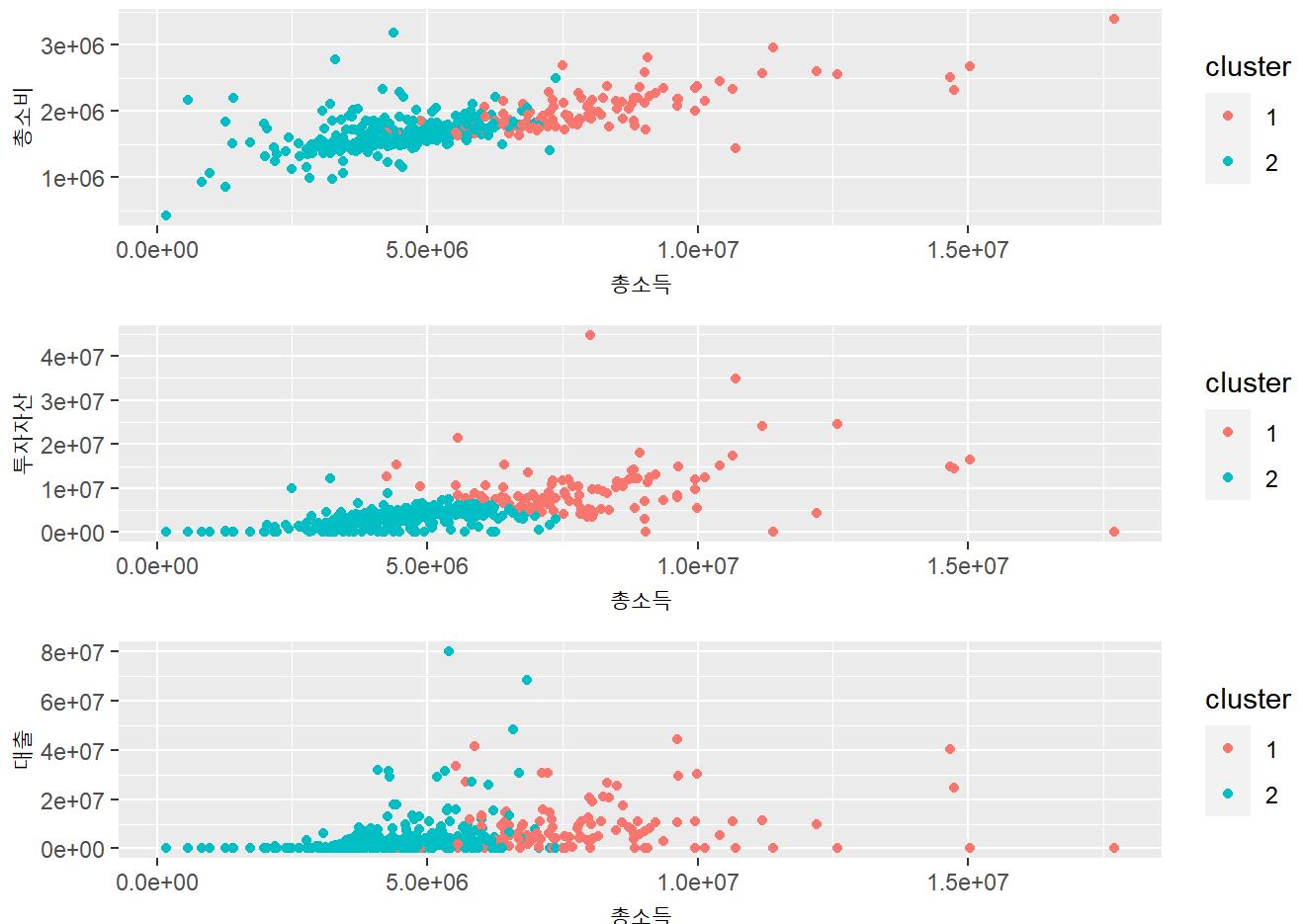
```
mean(총소득), 평균총소비 = mean(총소비), 평균투자자산 = mean(투자자산), 평균대출
= mean(대출))
```

```
## # A tibble: 2 x 5
##   cluster 평균총소득 평균총소비 평균투자자산 평균대출
##   <fct>     <dbl>     <dbl>     <dbl>     <dbl>
## 1 1          8027025.  2015188.  9415225.  8974270.
## 2 2          4380052.  1624940.  3194088.  3761980.
```

- 1.3 군집 시각화

```
g1 <- rst_o_2 %>% ggplot() + geom_point(aes(총소득, 총소비, col =
cluster))
g2 <- rst_o_2 %>% ggplot() + geom_point(aes(총소득, 투자자산, col =
cluster))
g3 <- rst_o_2 %>% ggplot() + geom_point(aes(총소득, 대출, col =
cluster))
```

```
grid.arrange(g1, g2, g3, ncol=1)
```



- 1.3 타깃 군집 특성 분석

```
## # A tibble: 25 x 2
##   지역구      n
##   <chr>     <int>
```

```

## 1 종로구      53
## 2 중구        40
## 3 성북구      34
## 4 영등포구    24
## 5 서대문구    17
## 6 용산구      17
## 7 마포구      15
## 8 성동구      13
## 9 은평구      11
## 10 강서구     10
## # ... with 15 more rows
## # A tibble: 314 x 2
## # Groups:   법정동 [314]
##   법정동     n
##   <chr>     <int>
## 1 봉천동      2
## 2 신대방동    2
## 3 신사동      2
## 4 신정동      2
## 5 홍은동      2
## 6 가락동      1
## 7 가리봉동    1
## 8 가산동      1
## 9 가양동      1
## 10 가회동     1
## # ... with 304 more rows
## # A tibble: 6 x 2
## # Groups:   나이 [6]
##   나이     n
##   <dbl> <int>
## 1 2 22339
## 2 3 32192
## 3 4 31699
## 4 5 31303
## 5 6 26509
## 6 7 16720
## # A tibble: 6 x 1
##   age_ratio
##   <dbl>
## 1 0.0714
## 2 0.0694
## 3 0.0723
## 4 0.0749
## 5 0.0740
## 6 0.0827

```

구별 분포를 살펴보면 종로구, 중구, 성북구, 영등포구에 많은 분포를 보이는 반면 동별 차이는 거의 없음. 타깃이 많은 분포를 보이는 3개의 지역구에서는 70대 노인이 전체 인구 집단에서 가장 큰 비중을 차지함.

## 2. 정보화 계수를 활용한 추가 분석

- 2.1 인구 데이터 불러오기
- 2.2 데이터 조인 및 정제

```
정보화 <- left_join(주민등록인구, 기초생활)
```

```
정보화 <- left_join(정보화, 장애인)
```

```
정보화 <- left_join(정보화, 고령자)
```

```
정보화$주민등록계 <- gsub(pattern=' ', replacement='', 정보화$주민등록계)
```

```
정보화$기초생활계 <- gsub(pattern=' ', replacement='', 정보화$기초생활계)
```

```
정보화$장애인계 <- gsub(pattern=' ', replacement='', 정보화$장애인계)
```

```
정보화$고령자계 <- gsub(pattern=' ', replacement='', 정보화$고령자계)
```

```
정보화$행정동 = as.factor(정보화$행정동)
```

```
정보화$주민등록계 = as.numeric(정보화$주민등록계)
```

```
정보화$기초생활계 = as.numeric(정보화$기초생활계)
```

```
정보화$장애인계 = as.numeric(정보화$장애인계)
```

```
정보화$고령자계 = as.numeric(정보화$고령자계)
```

- 2.2 정보화 계수 산출

```
정보화$정보화계수 <- (정보화$주민등록계 * 1 + 정보화$기초생활계*0.853 +  
    정보화$장애인계*0.668 + 정보화$고령자계*0.5) * 100
```

```
정보화 %>% as.tibble() %>% arrange(정보화계수)
```

```
## # A tibble: 439 x 6  
##   행정동     주민등록계  기초생활계  장애인계  고령자계  정보화계수  
##   <fct>       <dbl>        <dbl>        <dbl>        <dbl>        <dbl>  
## 1 을지로동      1869         125          87        462      226474.  
## 2 삼청동        2780         31          110        632      319592.  
## 3 소공동        3225         46          29        286      342661  
## 4 명동         3666         76          130        744      418967.  
## 5 가회동        4305         55          177        883      491165.  
## 6 필동          4729         76          173        890      535439.  
## 7 장충동        5310         99          198        830      594171.  
## 8 종로5·6가동    5554        247          255        1073     647153.  
## 9 회현동        5466        466          296        1340     673123.  
## 10 창신1동       5761        397          316        1229     692523.  
## # ... with 429 more rows
```

- 2.3 결과 데이터 내보내기

```
#동별 정보화 계수
```

```
#write.csv(정보화, 'C:/Users/Lee/Desktop/dat2.csv')
```

```
#target cluster 자산 정보
```

```
#write.csv(target_clu, 'C:/Users/Lee/Desktop/target.csv')
```

## 3. 타깃 집단의 정보화 계수 분석

- 3.1 데이터 불러오기

```
total<- read.csv('C:/Users/Lee/Desktop/tot2.csv', encoding =  
'UTF-8') [-1] %>% as_tibble()  
str(total)
```

```
total$법정동<- as.factor(total$법정동)
```

```
total %>% arrange(정보화계수)
sum(is.na(total))
```