2월 공공조달 빅데이터 경진대회 활동일지 – 상반기(5일 ~ 11일)

날짜	2월 5일	방법	Zoom 회의
활동	기업구분별 현황(기업수 분포 비율) 시각화 - 6개년	멤버	정진혁, 서지영

1-26. 기업구분별 현황 – Rstudio 코드

```r
# 필요한 세팅
rm(list = ls()); gc(reset = T)
require(dplyr); require(readxl)

# 엑셀 문서 읽기(1-26. 기업구분별 현황)
Excel_1 = read_xlsx("./조달청/1-26. 기업구분별 현황(2015_12~2016_11).xlsx") %>% data.frame(stringsAsFactors = F)
Excel_2 = read_xlsx("./조달청/1-26. 기업구분별 현황(2016_12~2017_11).xlsx") %>% data.frame(stringsAsFactors = F)
Excel_3 = read_xlsx("./조달청/1-26. 기업구분별 현황(2017_12~2018_11).xlsx") %>% data.frame(stringsAsFactors = F)
Excel_4 = read_xlsx("./조달청/1-26. 기업구분별 현황(2018_12~2019_11).xlsx") %>% data.frame(stringsAsFactors = F)
Excel_5 = read_xlsx("./조달청/1-26. 기업구분별 현황(2019_12~2020_11).xlsx") %>% data.frame(stringsAsFactors = F)
Excel_6 = read_xlsx("./조달청/1-26. 기업구분별 현황(2020_12~2021_11).xlsx") %>% data.frame(stringsAsFactors = F)

data_list = list(Excel_1, Excel_2, Excel_3, Excel_4, Excel_5, Excel_6)

# 중복제외한 기업수 연도별로 저장하기(final_year_num)
company = c("합계","대기업", "중견기업", "중소기업", "외국기업", "비영리법인 등 기타", "미분류")
year_num_list = list()
for(i in 1:6){
  year_num_list[[i]] = data_list[[i]][3, -1]
  year_num_list[[i]] = t(year_num_list[[i]])
  rownames(year_num_list[[i]]) = company
  colnames(year_num_list[[i]]) = 2015+i
}
final_year_num = year_num_list %>% do.call(cbind,) %>% data.frame(stringsAsFactors = F)
final_year_num = final_year_num %>% apply(., 2, as.numeric)

# 기업 퍼센트 비율 구하기(final_year_per)
final_year_per = final_year_num
for(i in 1:6){
  for(j in 1:7){
    final_year_per[j, i] = final_year_per[j, i]/final_year_num[1, i]*100
  }
}
final_year_per
rownames(final_year_per) = company
colnames(final_year_per) = c(2016:2021)
final_year_per = final_year_per %>% as.data.frame() %>% slice(c(2:nrow(final_year_per)))
final_year_per
```

```
> final_year_per
                  2016        2017        2018        2019        2020        2021
대기업          0.31480611  0.11394633  0.1128337  0.27295808  0.270223727  0.258112912
중견기업        0.59009110  0.30998054  0.2883035  0.34039479  0.416072252  0.417215325
중소기업       92.68178009 43.52284857 43.4329883 59.29673153 66.859643118 63.047708542
외국기업        0.05383048  0.01883603  0.0131047  0.01023593  0.004889965  0.004234665
비영리법인 등 기타  2.48097195  1.17806556  1.2305095  1.91632631  1.892416514  1.723710292
미분류          3.87852028 54.85632298 54.9222602 38.16335337 30.556754424 34.549018264
```

1-26. 기업구분별 현황 – Rstudio 코드

```r
# 시각화 – 년도별 기업수 비율(pie chart)(2016년 ~ 2021년)[6개년]
final_plot_2016 = final_year_per[,1] %>% as.data.frame()
rownames(final_plot_2016) = rownames(final_year_per)
final_plot_2016 = final_plot_2016 %>% mutate("기업구분" = rownames(final_plot_2016))
final_plot_2016 = final_plot_2016 %>% rename(value = ".")

plot_2016= ggplot(data = final_plot_2016) + geom_bar(aes(x = "", y = value, fill = as.factor(기업구분)), stat =
"identity") + coord_polar("y", start = 0) + labs(title = "2016년 기업구분별 현황")

final_plot_2017 = final_year_per[,2] %>% as.data.frame()
rownames(final_plot_2017) = rownames(final_year_per)
final_plot_2017 = final_plot_2017 %>% mutate("기업구분" = rownames(final_plot_2017))
final_plot_2017 = final_plot_2017 %>% rename(value = ".")

plot_2017= ggplot(data = final_plot_2017) + geom_bar(aes(x = "", y = value, fill = as.factor(기업구분)), stat =
"identity") + coord_polar("y", start = 0) + labs(title = "2017년 기업구분별 현황")

final_plot_2018 = final_year_per[,3] %>% as.data.frame()
rownames(final_plot_2018) = rownames(final_year_per)
final_plot_2018 = final_plot_2018 %>% mutate("기업구분" = rownames(final_plot_2017))
final_plot_2018 = final_plot_2018 %>% rename(value = ".")

plot_2018= ggplot(data = final_plot_2018) + geom_bar(aes(x = "", y = value, fill = as.factor(기업구분)), stat =
"identity") + coord_polar("y", start = 0) + labs(title = "2018년 기업구분별 현황")

final_plot_2019 = final_year_per[,4] %>% as.data.frame()
rownames(final_plot_2019) = rownames(final_year_per)
final_plot_2019 = final_plot_2019 %>% mutate("기업구분" = rownames(final_plot_2017))
final_plot_2019 = final_plot_2019 %>% rename(value = ".")

plot_2019= ggplot(data = final_plot_2019) + geom_bar(aes(x = "", y = value, fill = as.factor(기업구분)), stat =
"identity") + coord_polar("y", start = 0) + labs(title = "2019년 기업구분별 현황")

final_plot_2020 = final_year_per[,5] %>% as.data.frame()
rownames(final_plot_2020)= rownames(final_year_per)
final_plot_2020 = final_plot_2020 %>% mutate("기업구분" = rownames(final_plot_2017))
final_plot_2020 = final_plot_2020 %>% rename(value = ".")

plot_2020= ggplot(data = final_plot_2020) + geom_bar(aes(x = "", y = value, fill = as.factor(기업구분)), stat =
"identity") + coord_polar("y", start = 0) + labs(title = "2020년 기업구분별 현황")

final_plot_2021 = final_year_per[,6] %>% as.data.frame()
rownames(final_plot_2021)= rownames(final_year_per)
final_plot_2021 = final_plot_2021 %>% mutate("기업구분" = rownames(final_plot_2017))
final_plot_2021 = final_plot_2021 %>% rename(value = ".")

plot_2021= ggplot(data = final_plot_2021) + geom_bar(aes(x = "", y = value, fill = as.factor(기업구분)), stat =
"identity") + coord_polar("y", start = 0) + labs(title = "2021년 기업구분별 현황")
```

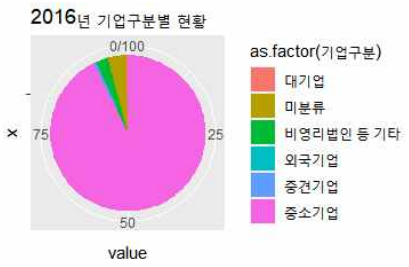
2017년을 기점으로 하여 중소기업의 절반 정도가 미분류로 바뀌었다. 외국기업은 매년 거의 존재하지 않는다고 생각해도 무방할 정도이며 대기업은 0.2%, 중견기업은 0.4%, 비영리법인은 1.7% 정도이다. 중소기업은 60%대를 유지하고 있다.

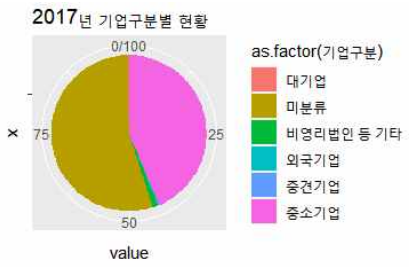
기업수의 분포는 이러한데 실제 기업 실적의 비율은 어떠할지 시각화를 통해 알아보고 불일치의 정도를 카이제곱 적합도 검정을 통해 알아볼 예정이다. 카이제곱 수치가 높을수록 불일치가 심하다고 결론 내릴 수 있다.

1-26. 기업구분별 현황(2015_12~2016_11).xlsx
1-26. 기업구분별 현황(2016_12~2017_11).xlsx
1-26. 기업구분별 현황(2017_12~2018_11).xlsx
1-26. 기업구분별 현황(2018_12~2019_11).xlsx
1-26. 기업구분별 현황(2019_12~2020_11).xlsx
1-26. 기업구분별 현황(2020_12~2021_11).xlsx

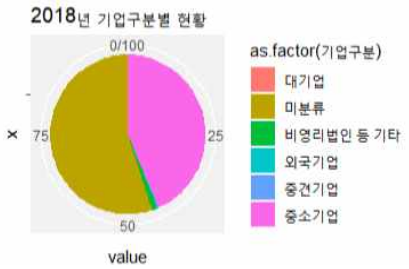
2월 공공조달 빅데이터 경진대회 활동일지 – 상반기(5일 ~ 11일)

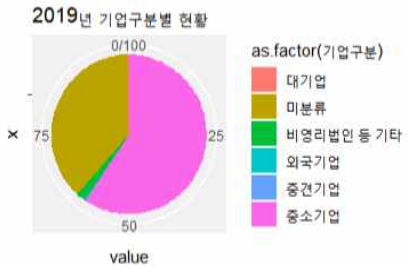
날짜	2월 5일	방법	Zoom 회의
활동	기업구분별 현황(기업수 분포 비율) 시각화 - 6개년	멤버	정진혁, 서지영

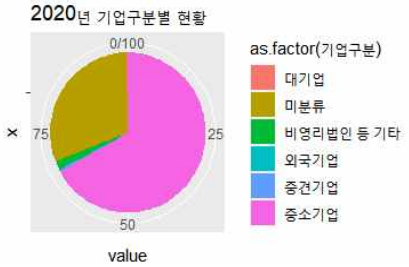


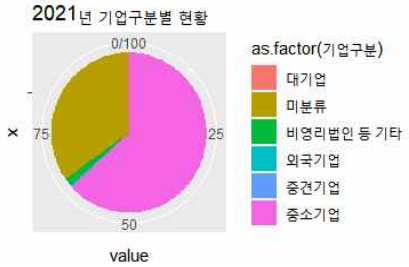



```
> final_year_per
                    2016        2017       2018        2019         2020         2021
대기업          0.31480611  0.11394633  0.1128337  0.27295808  0.270223727  0.258112912
중견기업        0.59009110  0.30998054  0.2883035  0.34039479  0.416072252  0.417215325
중소기업       92.68178009 43.52284857 43.4329883 59.29673153 66.859643118 63.047708542
외국기업        0.05383048  0.01883603  0.0131047  0.01023593  0.004889965  0.004234665
비영리법인 등 기타 2.48097195  1.17806556  1.2305095  1.91632631  1.892416514  1.723710292
미분류          3.87852028 54.85632298 54.9222602 38.16335337 30.556754424 34.549018264
```









< 기업 분류 기준 >

[대기업]

그 기업 집단에 소속된 회사의 자산 총액이 5조 이상 -> 공시대상기업집단으로 지정됨

공시대상기업집단 중 자산이 10조 이상되는 기업을 분류하여 상호출자제한기업으로 지정

[중견기업] → 아래 4가지 요건 중 하나만 만족해도 ok

1. 상시 직원 수 1000명 이상 / 2. 자산총액 5000억 이상

3. 자기자본 1000억 이상 / 4. 3년 평균 매출 1500억원 이상

[중소기업]

매출액 기준 : 평균 매출액 이하 / 자산총액 기준 : 5천억 미만 → 이 2가지 요건을 모두 만족

1-26. Rstudio 코드.hwp

2월 공공조달 빅데이터 경진대회 활동일지 – 상반기(5일 ~ 11일)

날짜	2월 6일	방법	Homework
활동	기업 조달 실적(건수, 금액 정리) + 금액 비율 시각화	멤버	정진혁, 서지영

1-3. 기업구분별 조달 현황 – Rstudio 코드

```
# 필요한 세팅
rm(list = ls()); gc(reset = T)
require(dplyr); require(readxl)

# 엑셀 데이터 읽기(기업구분별 조달 현황)
Excel_1 = read_xlsx("./조달청/1-3. 기업구분별 조달 현황(2015_12~2016_11).xlsx") %>% data.frame(stringsAsFactors = F)
Excel_2 = read_xlsx("./조달청/1-3. 기업구분별 조달 현황(2016_12~2017_11).xlsx") %>% data.frame(stringsAsFactors = F)
Excel_3 = read_xlsx("./조달청/1-3. 기업구분별 조달 현황(2017_12~2018_11).xlsx") %>% data.frame(stringsAsFactors = F)
Excel_4 = read_xlsx("./조달청/1-3. 기업구분별 조달 현황(2018_12~2019_11).xlsx") %>% data.frame(stringsAsFactors = F)
Excel_5 = read_xlsx("./조달청/1-3. 기업구분별 조달 현황(2019_12~2020_11).xlsx") %>% data.frame(stringsAsFactors = F)
Excel_6 = read_xlsx("./조달청/1-3. 기업구분별 조달 현황(2020_12~2021_11).xlsx") %>% data.frame(stringsAsFactors = F)

data_list = list(Excel_1, Excel_2, Excel_3, Excel_4, Excel_5, Excel_6)

# 2016년 ~ 2021년 기업별 조달 건수 데이터(year_num_list)
company = c("합계","대기업", "중견기업", "중소기업", "외국기업", "비영리법인 등 기타", "미분류")

year_num_list = list()
for(i in 1:6){
  year_num_list[[i]] = data_list[[i]][2, c(3, 5, 7, 9, 11, 13, 15)]
  colnames(year_num_list[[i]]) = company
  year_num_list[[i]] = t(year_num_list[[i]])
}
year_num_list = year_num_list %>% do.call(cbind,.)
year_num_list = year_num_list %>% apply(., 2, as.numeric) %>% data.frame(stringsAsFactors = F)
colnames(year_num_list) = c(2016:2021)
rownames(year_num_list) = company
year_num_list

# 2016년 ~ 2021년 기업별 조달 건수 퍼센트 데이터(year_num_per)
year_num_per = year_num_list
for(i in 1:6){
  for(j in 1:7){
    year_num_per[j, i] = year_num_per[j, i]/year_num_list[1, i]*100
  }
}
rownames(year_num_per) = company
year_num_per
```

```
[ 2016년 ~ 2021년 기업별 조달 실적 건수 및 비율 ]
> year_num_list
              2016     2017     2018     2019     2020     2021
합계       3221677  7095350  7563851  7980269  8345988  8615643
대기업      800813  977645  1020083  1381506  1738691  1292095
중견기업     46096   374443   483889   273170   286072  1079684
중소기업   2027991  3925805  4128657  5278927  5421855  6180109
외국기업       878     2706     2542     2687     2222     2990
비영리법인 등 기타  210398   310689   351524   452979   401078   432382
미분류      135301  1504062  1577156   591000   494870   618383
> year_num_per
                 2016      2017      2018      2019      2020      2021
합계       100.00000 100.00000 100.00000 100.00000 100.00000 100.00000
대기업      24.85702  13.77867  13.48629  17.31152  20.84703  13.43742
중견기업      1.43060   5.27730   6.39738   3.42306   3.42765  11.22841
중소기업     62.94830  55.32926  54.58406  66.14973  64.96360  64.37540
외국기업      0.02725   0.03817   0.03360   0.03367   0.02662   0.03109
비영리법인 등 기타  6.53069   4.37876   4.64742   5.67623   4.80565   4.49665
미분류      4.20591  21.19785  20.85123   7.40576   5.92943   6.43101
```

1-3. 기업구분별 조달 현황 – Rstudio 코드

```
# 2016년 ~ 2021년 기업별 조달 금액 데이터(year_price_list)
year_price_list = list()
for(i in 1:6){
  year_price_list[[i]] = data_list[[i]][2, c(4, 6, 8, 10, 12, 14, 16)]
  colnames(year_price_list[[i]]) = company
  year_price_list[[i]] = t(year_price_list[[i]])
}
year_price_list = year_price_list %>% do.call(cbind,.)
year_price_list = year_price_list %>% apply(., 2, as.numeric) %>% data.frame(stringsAsFactors = F)
colnames(year_price_list) = c(2016:2021)
rownames(year_price_list) = company
year_price_list

# 2016년 ~ 2021년 기업별 조달 금액 퍼센트 데이터(year_price_per)
year_price_per = year_price_list
for(i in 1:6){
  for(j in 1:7){
    year_price_per[j, i] = year_price_per[j, i]/year_price_list[1, i]*100
  }
}
rownames(year_price_per) = company
year_price_per

# 건수 비율과 금액 비율에서 함께 나타내는 행 제거하기(year_num_per, year_price_per)
year_num_per = year_num_per %>% as.data.frame() %>% slice(c(2:nrow(year_num_per)))
year_price_per = year_price_per %>% as.data.frame() %>% slice(c(2:nrow(year_price_per)))

year_num_per; year_price_per
```

```
> year_num_per
                 2016      2017      2018      2019      2020      2021
대기업      24.85702  13.77867  13.48629  17.31152  20.84703  13.43742
중견기업     1.43060   5.27730   6.39738   3.42306   3.42765  11.22841
중소기업    62.94830  55.32926  54.58406  66.14973  64.96360  64.37540
외국기업     0.02725   0.03817   0.03360   0.03367   0.02662   0.03109
비영리법인 등 기타  6.53069   4.37876   4.64742   5.67623   4.80565   4.49665
미분류      4.20591  21.19785  20.85123   7.40576   5.92943   6.43101
> year_price_per
                 2016      2017      2018      2019      2020      2021
대기업      15.9188874 13.3699008 12.4610613 12.1733346 11.2290096 12.8552555
중견기업    11.3491850 12.7242478 13.3948556 11.4615795 13.9562652 14.2447732
중소기업    64.4311965 64.1807125 64.6883273 66.1141891 66.1434866 64.1540261
외국기업     0.4795293  0.9013944  0.9220502  0.9474240  0.9748406  1.1569222
비영리법인 등 기타  7.4555818  7.3207118  6.5439485  8.3277318  6.8378775  6.8184844
미분류      0.3646200  1.5229928  1.9917571  0.9757409  0.7585204  0.7705386
```

```
[ 2016년 ~ 2021년 기업별 조달 실적 금액 및 비율 ]
> year_price_list
              2016     2017     2018     2019     2020     2021
합계       100.00000 100.00000 100.00000 100.00000 100.00000 100.00000
대기업      15.92887  13.36990  12.46061  12.17335  11.22901  12.85255
중견기업    11.34918  12.72424  13.39486  11.46157  13.95627  14.24477
중소기업    64.43119  64.18071  64.68827  66.11419  66.14348  64.15403
외국기업     0.47953   0.90139   0.92205   0.94742   0.97484   1.15692
비영리법인 등 기타  7.45582   7.32071  6.54395  8.32773  6.83785  6.81848
미분류      0.36462   1.52299  1.99175  0.97574  0.75852  0.77054
> year_price_per
                 2016      2017      2018      2019      2020      2021
합계       1.139918e+14 1.354489e+14 1.402734e+14 1.530087e+14 1.719941e+14
대기업      1.814776e+13 1.812998e+13 1.747723e+13 1.671359e+13 1.934168e+13
중견기업    1.293651e+13 1.723484e+13 1.878689e+13 1.753358e+13 2.399322e+13
중소기업    7.341987e+13 8.688506e+13 9.072825e+13 1.016845e+14 1.138869e+14
외국기업    5.467326e+11 1.220329e+12 1.293220e+12 1.457123e+12 1.675829e+12
비영리법인 등 기타  8.498025e+12 9.915891e+12 9.175776e+12 1.280787e+13 1.180512e+13
미분류      4.160596e+11 2.062878e+12 2.793525e+12 1.500776e+12 1.300559e+12
> year_price_per
                 2016      2017      2018      2019      2020      2021
합계       1.139918e+14
대기업      2.066157e+13
중견기업    2.610852e+13
중소기업    1.175688e+14
외국기업    2.129464e+12
비영리법인 등 기타  1.249726e+13
미분류      1.410779e+12
```

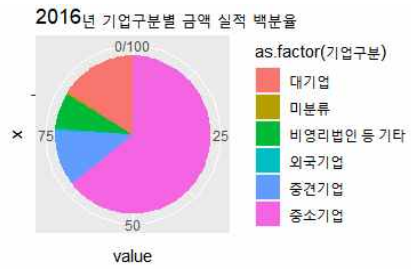
1-3. 기업구분별 조달 현황 – Rstudio 코드

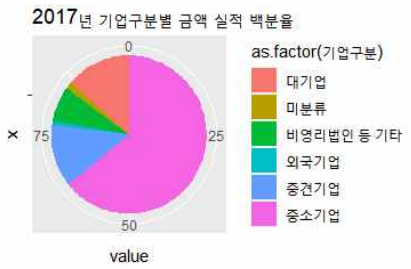
```
# 시각화 – 년도별 조달기업 실적 금액 비율(pie chart)
year_price_2016 = year_price_per[,1] %>% as.data.frame()
rownames(year_price_2016) = rownames(year_price_per)
year_price_2016 = year_price_2016 %>% rename(value = ".")
year_price_2016 = year_price_2016 %>% mutate("기업구분" = rownames(year_price_per))

plot_2016 = ggplot(data = year_price_2016) + geom_bar(aes(x = "", y = value, fill = as.factor(기업구분)), stat
= "identity") + coord_polar("y", start = 0) + labs(title = "2016년 기업구분별 금액 실적 백분율")

year_price_2017 = year_price_per[,2] %>% as.data.frame()
rownames(year_price_2017) = rownames(year_price_per)
year_price_2017 = year_price_2017 %>% rename(value = ".")
year_price_2017 = year_price_2017 %>% mutate("기업구분" = rownames(year_price_per))

plot_2017 = ggplot(data = year_price_2017) + geom_bar(aes(x = "", y = value, fill = as.factor(기업구분)), stat
= "identity") + coord_polar("y", start = 0) + labs(title = "2017년 기업구분별 금액 실적 백분율")

year_price_2018 = year_price_per[,3] %>% as.data.frame()
rownames(year_price_2018) = rownames(year_price_per)
year_price_2018 = year_price_2018 %>% rename(value = ".")
year_price_2018 = year_price_2018 %>% mutate("기업구분" = rownames(year_price_per))

plot_2018 = ggplot(data = year_price_2018) + geom_bar(aes(x = "", y = value, fill = as.factor(기업구분)), stat
= "identity") + coord_polar("y", start = 0) + labs(title = "2018년 기업구분별 금액 실적 백분율")

year_price_2019 = year_price_per[,4] %>% as.data.frame()
rownames(year_price_2019) = rownames(year_price_per)
year_price_2019 = year_price_2019 %>% rename(value = ".")
year_price_2019 = year_price_2019 %>% mutate("기업구분" = rownames(year_price_per))

plot_2019= ggplot(data = year_price_2019) + geom_bar(aes(x = "", y = value, fill = as.factor(기업구분)), stat
= "identity") + coord_polar("y", start = 0) + labs(title = "2019년 기업구분별 금액 실적 백분율")

year_price_2020 = year_price_per[,5] %>% as.data.frame()
rownames(year_price_2020) = rownames(year_price_per)
year_price_2020 = year_price_2020 %>% rename(value = ".")
year_price_2020 = year_price_2020 %>% mutate("기업구분" = rownames(year_price_per))

plot_2020= ggplot(data = year_price_2020) + geom_bar(aes(x = "", y = value, fill = as.factor(기업구분)), stat
= "identity") + coord_polar("y", start = 0) + labs(title = "2020년 기업구분별 금액 실적 백분율")

year_price_2021 = year_price_per[,6] %>% as.data.frame()
rownames(year_price_2021) = rownames(year_price_per)
year_price_2021 = year_price_2021 %>% rename(value = ".")
year_price_2021 = year_price_2021 %>% mutate("기업구분" = rownames(year_price_per))

plot_2021= ggplot(data = year_price_2021) + geom_bar(aes(x = "", y = value, fill = as.factor(기업구분)), stat
= "identity") + coord_polar("y", start = 0) + labs(title = "2021년 기업구분별 금액 실적 백분율")
```

1-3. 기업구분별 조달 현황(2015_12~2016_11).xlsx	1-3. 기업구분별 조달 현황(2018_12~2019_11).xlsx
1-3. 기업구분별 조달 현황(2016_12~2017_11).xlsx	1-3. 기업구분별 조달 현황(2019_12~2020_11).xlsx
1-3. 기업구분별 조달 현황(2017_12~2018_11).xlsx	1-3. 기업구분별 조달 현황(2020_12~2021_11).xlsx

2월 공공조달 빅데이터 경진대회 활동일지 – 상반기(5일 ~ 11일)

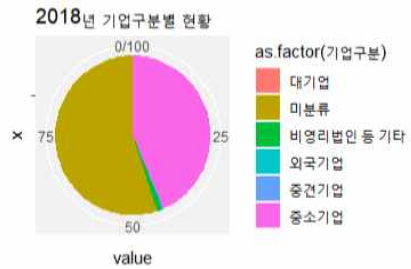
날짜	2월 6일	방법	Homework
활동	기업 조달 실적(건수, 금액 정리) + 금액 비율 시각화	멤버	정진혁

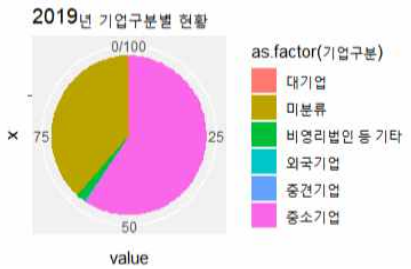




[기업별 조달 실적 금액 비율]

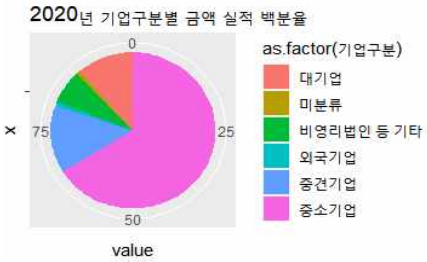
```
> year_price_per
                 2016       2017       2018       2019       2020       2021
대기업       15.9198874 13.3699008 12.4610613 12.1733346 11.2290096 12.8552555
중견기업     11.3491850 12.7242478 13.3948556 11.4615795 13.9562652 14.2447732
중소기업     64.4311965 64.1607125 64.6883273 66.1141891 66.1434866 64.1540261
외국기업      0.4795293  0.9013944  0.9220502  0.9474240  0.9748406  1.1569222
비영리법인 등 기타 7.4555818  7.3207518  6.5419485  8.3277319  6.9378775  6.8184844
미분류        0.3646200  1.5229928  1.9917571  0.9757409  0.7585204  0.7705386
```

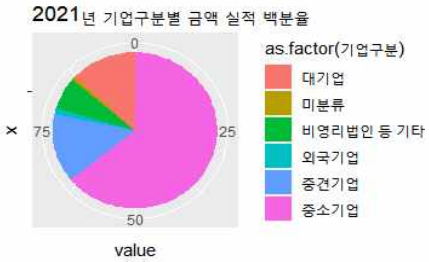




[기업수 분포 비율]

```
> final_year_per
                 2016        2017       2018        2019         2020         2021
대기업       0.31480611  0.11394633  0.1128337  0.27295808  0.270223727  0.258112912
중견기업     0.59009110  0.30998054  0.2883035  0.34039479  0.416072252  0.417215325
중소기업    92.68178009 43.52284857 43.4329883 59.29673153 66.859643118 63.047708542
외국기업     0.05383048  0.01883603  0.0131047  0.01023593  0.004889965  0.004234665
비영리법인 등 기타 2.48097195  1.17806556  1.2305095  1.91632631  1.892416514  1.723710292
미분류       3.87852028 54.85632298 54.9222602 38.16335337 30.556754424 34.549018264
```





[외국기업] - 외국 법인으로 분류된 기업

[비영리법인 등 기타]

비영리 법인 및 공공기관과 기업(단체), 자산총액 5조원 이상이지만 금융업 또는 보험업만 영위하는 기업, 중소기업과 상호출자제한기업에 속하지 않는 금융업, 보험 및 연금업, 금융 및 보험 관련 서비스업을 영위하는 기업

[미분류] - 나머지 5가지 기업에 모두 포함되지 않는 기업

1-3. Rstudio 코드.hwp

2월 공공조달 빅데이터 경진대회 활동일지 – 상반기(5일 ~ 11일)

날짜	2월 7일	방법	Homework
활동	기업구분별 현황(기업수 분포 비율) 시각화 - 6개년	멤버	정진혁

[기업별 조달 실적 금액 비율]

```
> year_price_per
```

	2016	2017	2018	2019	2020	2021
대기업	15.9198874	13.3699008	12.4610613	12.1733346	11.2290096	12.8552555
중견기업	11.3491850	12.7242478	13.3948556	11.4615795	13.9562652	14.2447732
중소기업	64.4311965	64.1607125	64.6883273	66.1141891	66.1434866	64.1540261
외국기업	0.4795293	0.9013944	0.9220502	0.9474240	0.9748406	1.1569222
비영리법인 등 기타	7.4555818	7.3207518	6.5419485	8.3277319	6.9378775	6.8184844
미분류	0.3646200	1.5229928	1.9917571	0.9757409	0.7585204	0.7705386

[기업수 분포 비율]

```
> final_year_per
```

	2016	2017	2018	2019	2020	2021
대기업	0.31480611	0.11394633	0.1128337	0.27295808	0.270223727	0.258112912
중견기업	0.59009110	0.30998054	0.2883035	0.34039479	0.416072252	0.417215325
중소기업	92.68178009	43.52284857	43.4329883	59.29673153	66.859643118	63.047708542
외국기업	0.05383048	0.01883603	0.0131047	0.01023593	0.004889965	0.004234665
비영리법인 등 기타	2.48097195	1.17806556	1.2305095	1.91632631	1.892416514	1.723710292
미분류	3.87852028	54.85632298	54.9222602	38.16335337	30.556754424	34.549018264

[편차 = 실제값(실적 금액 비율) - 평균(기업 분포 비율)]

```
> subtract_dat
```

	2016	2017	2018	2019	2020	2021
대기업	15.6050813	13.2559545	12.3482276	11.900377	10.9587859	12.597143
중견기업	10.7590939	12.4142672	13.1065521	11.121185	13.5401930	13.827558
중소기업	-28.2505836	20.6378639	21.2553390	6.817458	-0.7161565	1.106318
외국기업	0.4256988	0.8825584	0.9089455	0.937188	0.9699507	1.152688
비영리법인 등 기타	4.9746098	6.1426862	5.3114390	6.411406	5.0454610	5.094774
미분류	-3.5139002	-53.3333302	-52.9305031	-37.187613	-29.7982341	-33.778480

[대기업]

그 기업 집단에 소속된 회사의 자산 총액이 5조 이상 -> 공시대상기업집단으로 지정됨 공시대상기업집단 중 자산이 10조 이상되는 기업을 분류하여 상호출자제한기업으로 지정

[중견기업] → 아래 4가지 요건 중 하나만 만족해도 ok

2. 상시 직원 수 1000명 이상 / 2. 자산총액 5000억 이상

3. 자기자본 1000억 이상 / 4. 3년 평균 매출 1500억원 이상

[중소기업] → 이 2가지 요건을 모두 만족

매출액 기준 : 평균 매출액 이하 / 자산총액 기준 : 5천억 미만

[외국기업] - 외국 법인으로 분류된 기업

[비영리법인 등 기타]

비영리 법인 및 공공기관과 기업(단체), 자산총액 5조원 이상이지만 금융업 또는 보험업만 영위하는 기업, 중소기업과 상호출자제한기업에 속하지 않는 금융업, 보험 및 연금업, 금융 및 보험 관련 서비스업을 영위하는 기업

[미분류] - 나머지 5가지 기업에 모두 포함되지 않는 기업

비교적 자산 총액 규모가 큰 대기업과 중견기업의 경우 기업 분포 비율보다 실적 금액 비율이 상당히 높았다. 이는 편차를 통해서도 확인이 가능하다. 중소기업의 경우 시간이 갈수록 편차가 점점 줄어들고 있으며 외국 기업과 비영리 법인 등 기타 부문에서는 큰 차이가 없었다. 하지만 미분류 실적은 실적(금액 비율)이 저조하다는 것을 알 수 있었다.

비영리 법인이 아니므로 영리 목적을 추구하는 기업임을 알 수 있고 외국 기업은 아니다. 아마 매출액이 자산총액 기준으로는 5천억 미만이지만 평균 매출액 이상이라 미분류로 분류되었을 가능성이 높다.

1-3. Rstudio 코드.hwp

2월 공공조달 빅데이터 경진대회 활동일지 – 상반기(5일 ~ 11일)

날짜	2월 8일	방법	Homework
활동	카이제곱 적합도 검정 공부하기 / 적용(16~21)	멤버	정진혁

[카이제곱 적합도 검정]

변수가 지정된 분포에서 추출될 가능성이 있는지 여부를 확인하는 데 사용되는 통계적 가설 검정

대개 표본 데이터가 전체 모집단을 대표하는지 평가하는 데 사용된다.

범주형 변수 값의 개수가 있을 대 사용할 수 있는 검정이다.

[카이제곱 적합도 검정 예시]

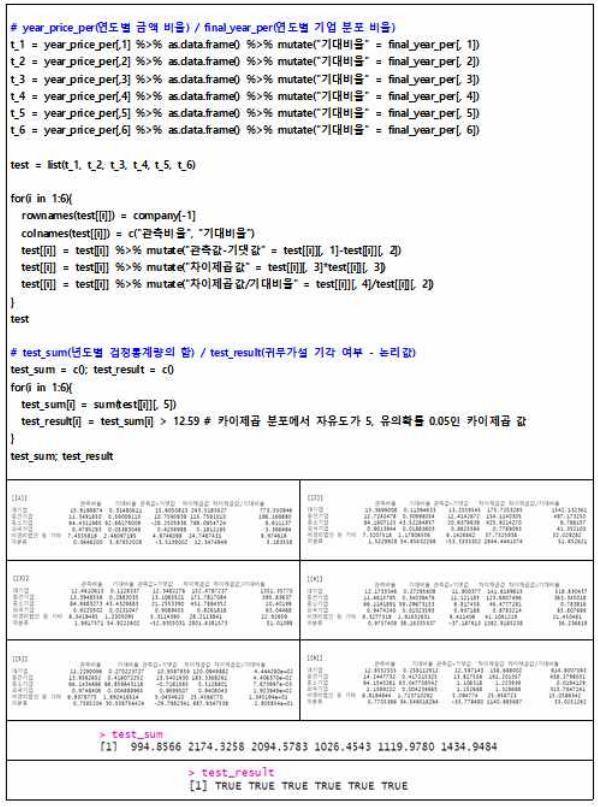
캔디맛	캔디 개수 (10봉지)	캔디 기대 개수	관측값-기대값	차이 제곱값	차이제곱값 /기대 개수
사과	180	200	-20	400	2
라임	250	200	50	2500	12.5
체리	120	200	-80	6400	32
오렌지	225	200	25	625	3.125
포도	225	200	25	625	3.125

유의확률이 0.05이고 자유도가 4인 카이제곱 값은 9.488

검정 통계량은 52.75

검정통계량이 카이제곱 값보다 크므로 캔디 맛별 비율이 동일하다는 귀무가설 기각

카이제곱 적합도 검정(기업분포별 금액) – Rstudio 코드



```
# year_price_per(연도별 금액 비율) / final_year_per(연도별 기업 분포 비율)
t_1 = year_price_per[,1] %>% as.data.frame() %>% mutate("기대비율" = final_year_per[, 1])
t_2 = year_price_per[,2] %>% as.data.frame() %>% mutate("기대비율" = final_year_per[, 2])
t_3 = year_price_per[,3] %>% as.data.frame() %>% mutate("기대비율" = final_year_per[, 3])
t_4 = year_price_per[,4] %>% as.data.frame() %>% mutate("기대비율" = final_year_per[, 4])
t_5 = year_price_per[,5] %>% as.data.frame() %>% mutate("기대비율" = final_year_per[, 5])
t_6 = year_price_per[,6] %>% as.data.frame() %>% mutate("기대비율" = final_year_per[, 6])

test = list(t_1, t_2, t_3, t_4, t_5, t_6)

for(i in 1:6){
  rownames(test[[i]]) = company[-1]
  colnames(test[[i]]) = c("관측비율", "기대비율")
  test[[i]] = test[[i]] %>% mutate("관측값-기댓값" = test[[i]][, 1]-test[[i]][, 2])
  test[[i]] = test[[i]] %>% mutate("차이제곱값" = test[[i]][, 3]*test[[i]][, 3])
  test[[i]] = test[[i]] %>% mutate("차이제곱값/기대비율" = test[[i]][, 4]/test[[i]][, 2])
}
test

# test_sum(년도별 검정통계량의 합) / test_result(귀무가설 기각 여부 - 논리값)
test_sum = c(); test_result = c()
for(i in 1:6){
  test_sum[i] = sum(test[[i]][, 5])
  test_result[i] = test_sum[i] > 12.59 # 카이제곱 분포에서 자유도가 5, 유의확률 0.05인 카이제곱 값
}
test_sum; test_result
```

```
> test_sum
[1]  994.8566 2174.3258 2094.5783 1026.4543 1119.9780 1434.9484

> test_result
[1] TRUE TRUE TRUE TRUE TRUE TRUE
```

1-26. 기업구분별 현황(2015_12~2016_11).xlsx	1-26. 기업구분별 현황(2018_12~2019_11).xlsx	1-3. 기업구분별 조달 현황(2015_12~2016_11).xlsx	1-3. 기업구분별 조달 현황(2018_12~2019_11).xlsx
1-26. 기업구분별 현황(2016_12~2017_11).xlsx	1-26. 기업구분별 현황(2019_12~2020_11).xlsx	1-3. 기업구분별 조달 현황(2016_12~2017_11).xlsx	1-3. 기업구분별 조달 현황(2019_12~2020_11).xlsx
1-26. 기업구분별 현황(2017_12~2018_11).xlsx	1-26. 기업구분별 현황(2020_12~2021_11).xlsx	1-3. 기업구분별 조달 현황(2017_12~2018_11).xlsx	1-3. 기업구분별 조달 현황(2020_12~2021_11).xlsx

https://www.jmp.com/ko_kr/statistics-knowledge-portal/chi-square-test/chi-square-goodness-of-fit-test.html

카이제곱 적합도 검정(기업분포별 금액). Rstudio 코드.hwp

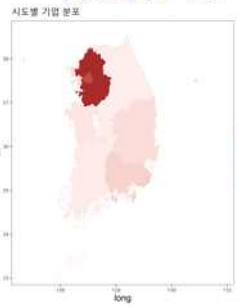
2월 공공조달 빅데이터 경진대회 활동일지 – 상반기(5일 ~ 11일)

날짜	2월 9일	방법	Homework
활동	지역별 기업수 분포 비율 시각화 - 4개년	멤버	서지영

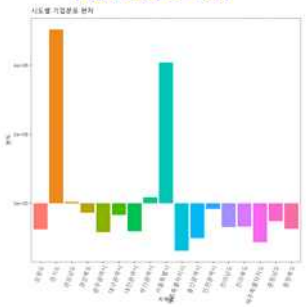
산업별 지역별 기업수 – Rstudio 코드

```
# 필요한 세팅
rm(list=ls()); gc(reset = T)
require(readxl); require(dplyr)
library(ggmap); library(ggplot2); library(raster)
library(rgeos); library(maptools); library(rgdal)

# 지역별 기업 수 데이터 읽기
Excel = read_xlsx("./산업별 지역별 기업수(2017~2020).xlsx") %>% data.frame(stringsAsFactors = F)
Excel = Excel[-1, -1]
colnames(Excel) = c("지역", 2017, 2018, 2019, 2020)

# 지역명 저장하기
region = Excel[, 1]

# 지역명 없애고 숫자형으로 바꾸기(계산하기 위해)
Excel = Excel[, -1] %>% apply(., 2, as.numeric) %>% as.data.frame()

Excel = Excel %>% mutate("기업수_합계" = c(1:nrow(Excel)))
for(i in 1:nrow(Excel)){
  Excel[i, 5] = Excel[i, 1] + Excel[i, 2] + Excel[i, 3] + Excel[i, 4]
}
Excel = Excel[, c(4, 5)]
Excel = Excel %>% mutate("지역명" = region)
Excel = Excel[, c(2, 3)]
Excel = Excel[, c(2, 1)]
Excel = Excel %>% slice(c(2:nrow(Excel))) # 전국 제외
```

[시도별 기업 분포 – 지도]



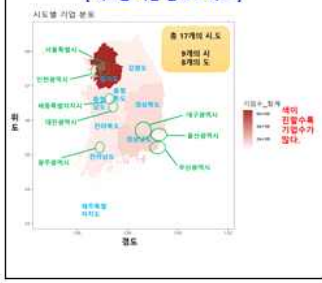
[시도별 기업 분포 – 편차]



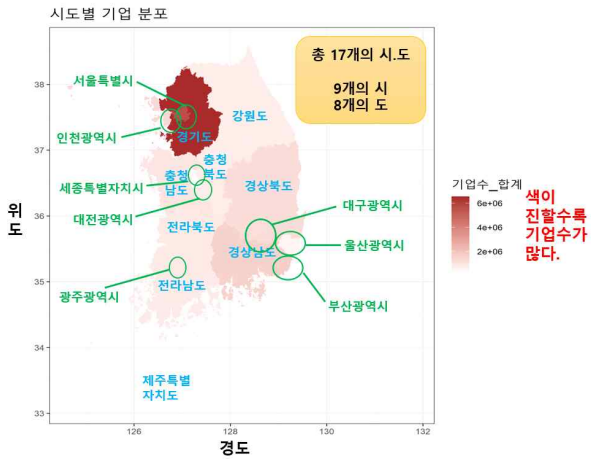
산업별 지역별 기업수 – Rstudio 코드

```
# 데이터 시각화 - 지도
Excel$id = 0:16

korea = shapefile("TL_SCCO_CTPRVN.shp")
korea = spTransform(korea, CRS("+proj=longlat"))
korea_map = fortify(korea)
merge_result = merge(korea_map, Excel, by = "id")

plot = ggplot(data = merge_result) + geom_polygon(aes(x = long, y = lat, group = group, fill = 기업수_합
계)) + labs(title = "시도별 기업 분포") + theme_bw() + theme(title = element_text(size = 15))

plot = plot + scale_fill_gradient2(low = "lightyellow", mid = "white", high = "brown", midpoint = .02)
plot


# 데이터 시각화 – 편차
Excel = Excel %>% mutate(편차 = 기업수_합계 - mean(기업수_합계))

sd_plot = ggplot(data = Excel) + geom_bar(aes(x = 지역명, y = 편차, fill = as.factor(지역명)), stat =
"identity") + theme(title = element_text(size = 15)) + theme_bw() + theme(axis.text.x = element_text(angle =
70, hjust = 1)) + labs(title = "시도별 기업분포 편차") + theme(legend.position = "none")

sd_plot
```

[시도별 기업 분포 – 지도]



[시도별 기업 분포 – 편차]



산업별 지역별 기업수(2017~2020).xlsx

산업별 지역별 기업수. Rstudio 코드.hwp

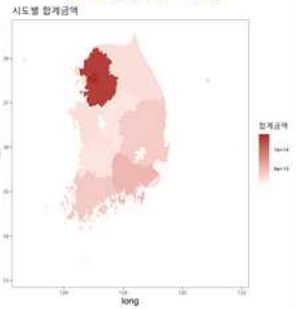
2월 공공조달 빅데이터 경진대회 활동일지 – 상반기(5일 ~ 11일)

날짜	2월 10일	방법	Homework
활동	지역별 기업 조달 실적 비율 시각화 - 4개년	멤버	서지영

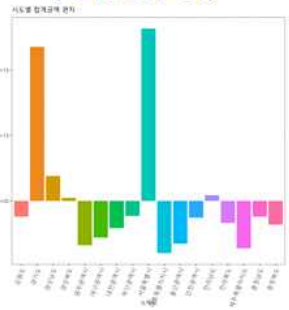
1-33. 조달기업별 실적 순위 – Rstudio 코드

```
# 필요한 세팅
rm(list=ls()); gc(reset = T)
require(readxl); require(dplyr)
library(ggmap); library(ggplot2); library(raster)
library(rgeos); library(maptools); library(rgdal)

# 금액과 건수로 나누어 데이터 읽기(price_dat, num_dat)
price_dat = read_xlsx("./1-33. 조달기업별 실적 순위(2017~2020).xlsx") %>% data.frame(stringsAsFactors = F)
price_dat = price_dat %>% slice(c(2:nrow(price_dat))) %>% filter(건수 > 0 & 금액 > 0)
num_dat = price_dat %>% arrange(건수)
price = price_dat$금액
head(price, 10)
tail(price, 10)

# 지역별 합계 금액 데이터
unique(price_dat$소재지)
local_dat = price_dat %>% group_by(., 소재지) %>% summarise_at(vars(금액), sum, na.rm = T) %>%
data.frame(stringsAsFactors = F) %>% rename(합계금액 = 금액)
local_dat = local_dat %>% arrange(합계금액)
local_dat = local_dat[-c(1,7), ] #미분류, 국외소재 제외(17개 시.도구분으로 바꿈)
```

[시도별 합계 금액 – 지도]



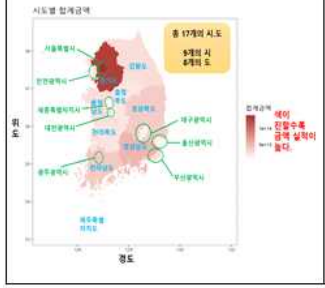
[시도별 합계 금액 – 편차]



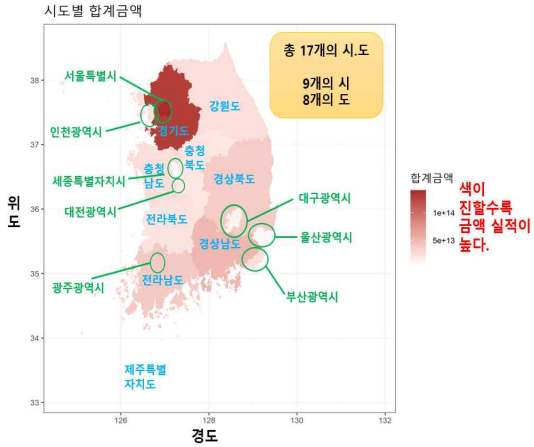
1-33. 조달기업별 실적 순위 – Rstudio 코드

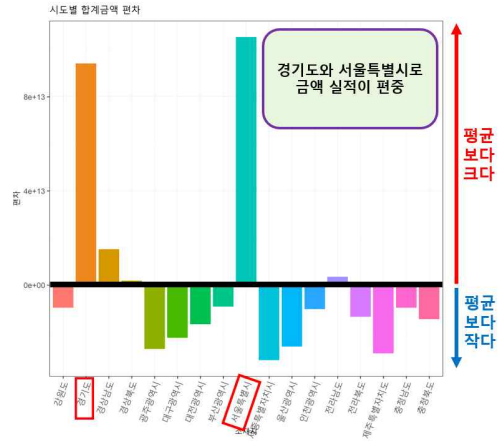
```
# 데이터 시각화 - 지도
local_dat = local_dat[c(17,12,5,9,3,6,4,1,16,11,7,10,8,14,13,15,2),]
rownames(local_dat) = c(1:nrow(local_dat))
local_dat$id = 0:16

korea = shapefile("TL_SCCO_CTPRVN.shp")
korea = spTransform(korea, CRS("+proj=longlat"))
korea_map = fortify(korea)
merge_result = merge(korea_map, local_dat, by = "id")

plot = ggplot(data = merge_result) + geom_polygon(aes(x = long, y = lat, group = group, fill = 합계금액))
+ labs(title = "시도별 합계금액") + theme_bw() + theme(title = element_text(size = 15))

plot = plot + scale_fill_gradient2(low = "lightyellow", mid = "white", high = "brown", midpoint = .02)
plot

# 데이터 시각화 - 편차
local_dat = local_dat %>% mutate(편차 = 합계금액 - mean(합계금액))

sd_plot = ggplot(data = local_dat) + geom_bar(aes(x = 소재지, y = 편차, fill = as.factor(소재지)), stat =
"identity") + theme(title = element_text(size = 15)) + theme_bw() + theme(axis.text.x = element_text(angle =
70, hjust = 1)) + labs(title = "시도별 합계금액 편차") + theme(legend.position = "none")

sd_plot
```

[시도별 합계 금액 – 지도]



[시도별 합계 금액 – 편차]





1-33. 조달기업별 실적 순위(2017~2020).xlsx

1-33. Rstudio 코드.hwp

2월 공공조달 빅데이터 경진대회 활동일지 – 상반기(5일 ~ 11일)

날짜	2월 11일	방법	Homework
활동	카이제곱 적합도 검정 적용 – 지역별 조달 실적(17~20)	멤버	정진혁

카이제곱 적합도 검정(지역별 실적 분포) – Rstudio 코드

```
# 필요한 세팅
rm(list = ls()); gc(reset = T)
require(readxl); require(dplyr)

# 지역별 기업 수 데이터 읽기
Excel = read_xlsx("./조달청/산업별 지역별 기업수(2017~2020).xlsx") %>% data.frame(stringsAsFactors = F)
Excel = Excel[-1, -1]
colnames(Excel) = c("지역", 2017, 2018, 2019, 2020)

# 지역명 저장하기
region = Excel[, 1]

# 지역명 없애고 숫자형으로 바꾸기(계산하기 위해)Excel : 기업수 2017~2020 기업수 비율 ]
Excel = Excel[, -1] %>% apply(., 2, as.numeric) %>% as.data.frame()
rownames(Excel) = region
Excel
sum_num = Excel[1, ]
Excel = Excel[-1, ]
for(i in 1:4){
  for(j in 1:nrow(Excel)){
    Excel[j, i] = Excel[j, i]/sum_num[i]*100
  }
}
per_dat = Excel
```

```
> Excel
                 2017       2018       2019       2020
서울특별시   22.1751114 21.9539747 21.6982372 21.4236349
부산광역시    6.6967809  6.6024842  6.5212615  6.4366318
대구광역시    4.6791444  4.6156475  4.5357237  4.4525096
인천광역시    5.1196192  5.1481299  5.3499371  5.3538049
광주광역시    2.5755019  2.6389392  2.6207745  2.5931519
대전광역시    2.7500590  2.7174539  2.6906583  2.6795781
울산광역시    2.0094722  1.9657479  1.9151637  1.8760712
세종특별자치시 0.4312859  0.4836167  0.5251784  0.5634928
경기도       24.8326032 25.3599324 25.7208869 26.1577956
강원도        2.9402174  2.9306773  2.9177157  2.9192109
충청북도      3.0148033  2.9831271  2.9873544  2.9933513
충청남도      3.8208035  3.8384207  3.8334990  3.8954822
전라북도      3.2389411  3.2163192  3.2477679  3.2954690
전라남도      3.1745523  3.1510849  3.1773324  3.2172676
경상북도      4.9138428  4.8367912  4.7716219  4.7248803
경상남도      6.1846124  6.0754100  5.9788151  5.9200393
제주특별자치도 1.4426493  1.4822432  1.5080723  1.4976286
```

카이제곱 적합도 검정(지역별 실적 분포) – Rstudio 코드

```
# 시작 전
region = region[-1]
per_group = list(); region_list = list(); sum_price = c()

# 2017년
Excel = read_xlsx("./조달청/1-33. 조달기업별 실적 순위(2017).xlsx") %>% data.frame(stringsAsFactors = F)
Excel = Excel %>% slice(c(2:nrow(Excel))) %>% filter("건수" > 0 & "X0" > 0)
Excel

for(i in 1:length(region)){
  region_list[[i]] = Excel[, c(2, 5)] %>% filter(소재지 == region[i])
  sum_price[i] = sum(as.numeric(region_list[[i]][, 2]))
}

final_dat = data.frame(합계금액 = sum_price)
rownames(final_dat) = region
final_dat = final_dat %>% apply(., 2, as.numeric) %>% as.data.frame()
final_dat = final_dat %>% mutate(합계금액비율 = 1:nrow(final_dat))

for(i in 1:nrow(final_dat)){
  final_dat[i, 2] = final_dat[i, 1]/sum(final_dat[, 1])*100
}
rownames(final_dat) = region
per_group[[1]] = final_dat

# 2018년
Excel = read_xlsx("./조달청/1-33. 조달기업별 실적 순위(2018).xlsx") %>% data.frame(stringsAsFactors = F)
Excel = Excel %>% slice(c(2:nrow(Excel))) %>% filter("건수" > 0 & "X0" > 0)
Excel

for(i in 1:length(region)){
  region_list[[i]] = Excel[, c(2, 5)] %>% filter(소재지 == region[i])
  sum_price[i] = sum(as.numeric(region_list[[i]][, 2]))
}

final_dat = data.frame(합계금액 = sum_price)
rownames(final_dat) = region
final_dat = final_dat %>% apply(., 2, as.numeric) %>% as.data.frame()
final_dat = final_dat %>% mutate(합계금액비율 = 1:nrow(final_dat))

for(i in 1:nrow(final_dat)){
  final_dat[i, 2] = final_dat[i, 1]/sum(final_dat[, 1])*100
}
rownames(final_dat) = region
per_group[[2]] = final_dat
```

카이제곱 적합도 검정(지역별 실적 분포) – Rstudio 코드

```
# 2019년
Excel = read_xlsx("./조달청/1-33. 조달기업별 실적 순위(2019).xlsx") %>% data.frame(stringsAsFactors = F)
Excel = Excel %>% slice(c(2:nrow(Excel))) %>% filter("건수" > 0 & "X0" > 0)
Excel

for(i in 1:length(region)){
  region_list[[i]] = Excel[, c(2, 5)] %>% filter(소재지 == region[i])
  sum_price[i] = sum(as.numeric(region_list[[i]][, 2]))
}

final_dat = data.frame(합계금액 = sum_price)
rownames(final_dat) = region
final_dat = final_dat %>% apply(., 2, as.numeric) %>% as.data.frame()
final_dat = final_dat %>% mutate(합계금액비율 = 1:nrow(final_dat))

for(i in 1:nrow(final_dat)){
  final_dat[i, 2] = final_dat[i, 1]/sum(final_dat[, 1])*100
}
rownames(final_dat) = region
per_group[[3]] = final_dat

# 2020년
Excel = read_xlsx("./조달청/1-33. 조달기업별 실적 순위(2020).xlsx") %>% data.frame(stringsAsFactors = F)
Excel = Excel %>% slice(c(2:nrow(Excel))) %>% filter("건수" > 0 & "X0" > 0)
Excel

for(i in 1:length(region)){
  region_list[[i]] = Excel[, c(2, 5)] %>% filter(소재지 == region[i])
  sum_price[i] = sum(as.numeric(region_list[[i]][, 2]))
}

final_dat = data.frame(합계금액 = sum_price)
rownames(final_dat) = region
final_dat = final_dat %>% apply(., 2, as.numeric) %>% as.data.frame()
final_dat = final_dat %>% mutate(합계금액비율 = 1:nrow(final_dat))

for(i in 1:nrow(final_dat)){
  final_dat[i, 2] = final_dat[i, 1]/sum(final_dat[, 1])*100
}
rownames(final_dat) = region
per_group[[4]] = final_dat
```

카이제곱 적합도 검정(지역별 실적 분포) – Rstudio 코드

```
test = list()
for(i in 1:4){
  test[[i]] = data.frame("금액비율" = per_group[[i]][, 2], "기대비율" = per_dat[[i]])
  test[[i]] = test[[i]] %>% mutate("관측값-기댓값" = test[[i]][, 1]-test[[i]][, 2])
  test[[i]] = test[[i]] %>% mutate("차이제곱값" = test[[i]][, 3]*test[[i]][, 3])
  test[[i]] = test[[i]] %>% mutate("차이제곱값/기대비율" = test[[i]][, 4]/test[[i]][, 2])
}
test

# test_sum : 검정통계량 / test_result : 귀무가설 기각 여부
test_sum = c(); test_result = c()
for(i in 1:4){
  test_sum[i] = sum(test[[i]][, 5])
  test_result[i] = test_sum[i] > 26.3 # 카이제곱 분포에서 자유도가 16, 유의확률 0.05인 카이제곱 값
}
test_sum; test_result
```

```
> test_sum
[1]  9.079502  9.817105 10.042422  9.696607

> test_result
[1] FALSE FALSE FALSE FALSE
```

산업별 지역별 기업수(2017~2020).xlsx
1-33. 조달기업별 실적 순위(2017).xlsx
1-33. 조달기업별 실적 순위(2018).xlsx

카이제곱 적합도 검정(지역별 실적 분포). Rstudio 코드.hwp
1-33. 조달기업별 실적 순위(2019).xlsx
1-33. 조달기업별 실적 순위(2020).xlsx