R 데이터(텍스트) 마이닝 및 시각화 분석

Seongmin Mun

15th January 2022

아주대학교 AJOU UNIVERSITY

Outline

분산분석(ANOVA, analysis of variance)

ggplot2를 이용한 시각화 분석

분산분석(ANOVA, analysis of variance)

분산분석(ANOVA)

- n개의 집단을 비교하는 통계적 분포 (단, n > 2)

분산분석(ANOVA)의 종류

- 일원배치 분산분석(one-way ANOVA)
  - 분산분석 중 종속변인이 1개, 독립변인이 1개인 경우 수행하는 분산분석
- 이원배치 분산분석(two-way ANOVA)
  - 분산분석 중 종속변인이 1개, 독립변인이 2개인 경우 수행하는 분산분석

네카라쿠배당토

데이터 생성

```
> #네카라쿠배당토
> coupang <- sample(c(6000:9000),100,replace = TRUE)
> kakao <- sample(c(4000:9000),100,replace = TRUE)
> naver <- sample(c(4000:8000),100,replace = TRUE)
> companyName <- c("coupang","kakao","naver")
> companies <- NULL
> income <- NULL
> employee <- NULL
> for(i in 1:3){
+   for (j in 1:100){
+     companies <- c(companies, i)
+     employee <- c(employee,j)
+     if (i==1){
+       income <- c(income, coupang[j])
+     } else if (i==2){
+       income <- c(income, kakao[j])
+     } else {
+       income <- c(income, naver[j])
+     }
+   }
+ }
```

데이터 확인

```
> overall<-data.frame(Company=companies,Employee=employee,Income=income)
> head(overall)
  Company Employee Income
1       1        1   8432
2       1        2   7394
3       1        3   6345
4       1        4   7799
5       1        5   6147
6       1        6   7624
> str(overall)
'data.frame':   300 obs. of  3 variables:
 $ Company : int  1 1 1 1 1 1 1 1 1 1 ...
 $ Employee: int  1 2 3 4 5 6 7 8 9 10 ...
 $ Income  : int  8432 7394 6345 7799 6147 7624 6315 8417 6168 6714 ...
> summary(overall)
    Company     Employee          Income
 Min.   :1   Min.   :  1.00   Min.   :4016
 1st Qu.:1   1st Qu.: 25.75   1st Qu.:5774
 Median :2   Median : 50.50   Median :6777
 Mean   :2   Mean   : 50.50   Mean   :6738
 3rd Qu.:3   3rd Qu.: 75.25   3rd Qu.:7811
 Max.   :3   Max.   :100.00   Max.   :8970
```

일원배치 분산분석(one-way ANOVA)

- H0: 세 회사간 연봉에서는 차이가 없음
- H1: 세 회사간 연봉에서는 차이가 있음

일원배치 분산분석(one-way ANOVA)

```
> aov.ex = aov(Income~Company,data=overall)
> summary(aov.ex)
             Df    Sum Sq  Mean Sq F value   Pr(>F)
Company       1  96191224 96191224   64.88 1.93e-14 ***
Residuals   298 441806835  1482573
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

이원배치 분산분석(two-way ANOVA)

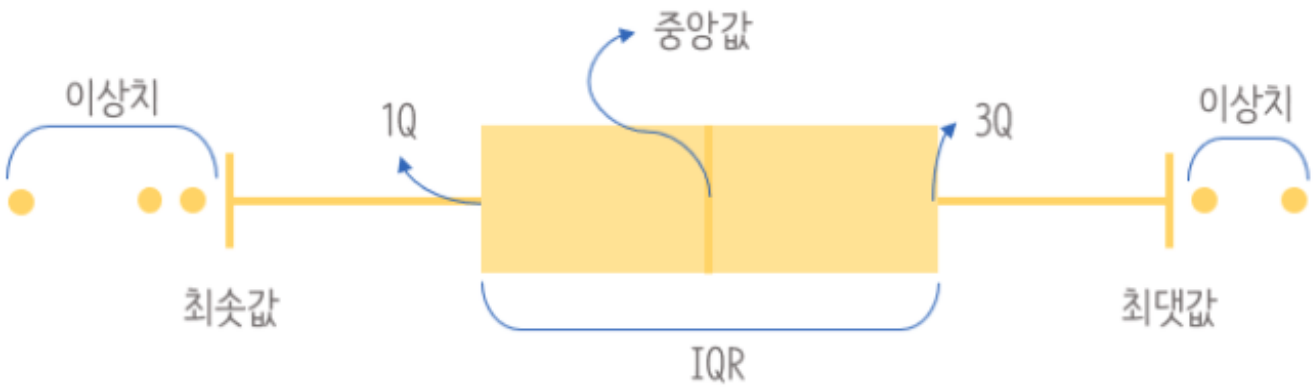
- 세 회사간 연봉의 차이
- 직원들간 연봉의 차이
- 각 회사 내에서 직원들간 연봉의 차이

이원배치 분산분석(two-way ANOVA)

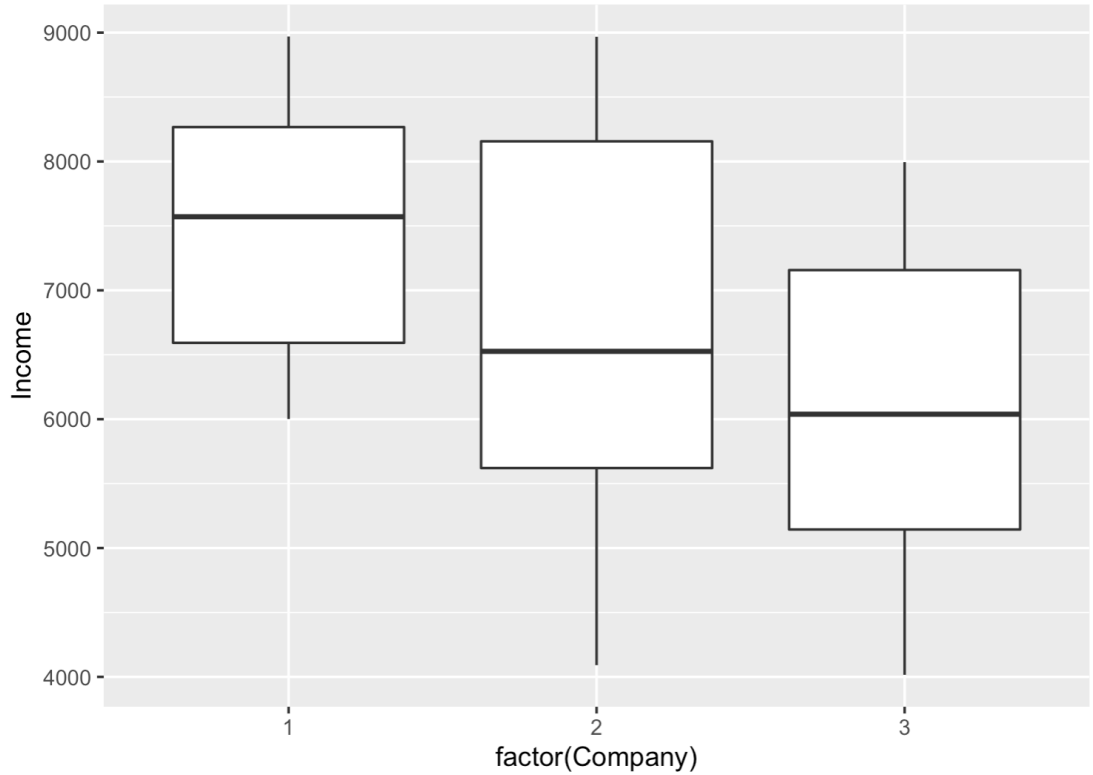
```
> aov.ex = aov(Income~Company*Employee,data=overall)
> summary(aov.ex)
                  Df    Sum Sq  Mean Sq F value   Pr(>F)
Company            1  96191224 96191224  64.642 2.17e-14 ***
Employee           1   1159410  1159410   0.779    0.378
Company:Employee   1    182860   182860   0.123    0.726
Residuals        296 440464565  1488056
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

ggplot2를 이용한 시각화 분석

상자그림



상자 그림 그리기

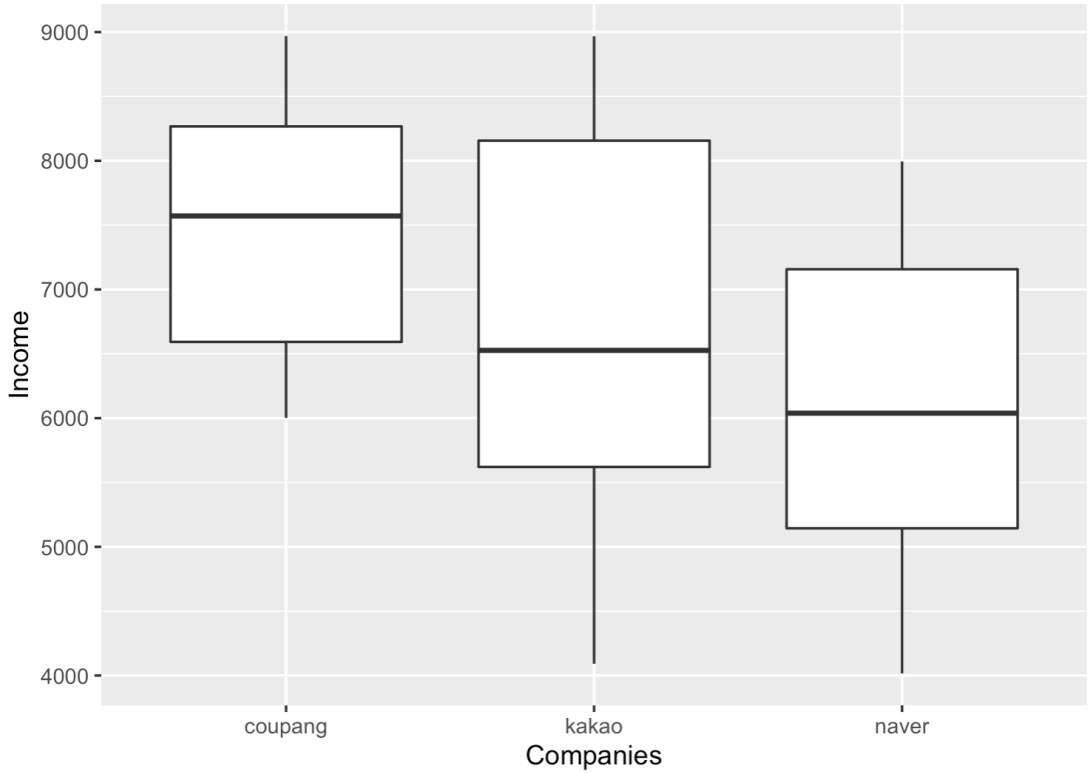



```
> library(ggplot2)
> ggplot(overall, aes(x=factor(Company),y=Income)) + geom_boxplot()
```

문자형 데이터 추가하기

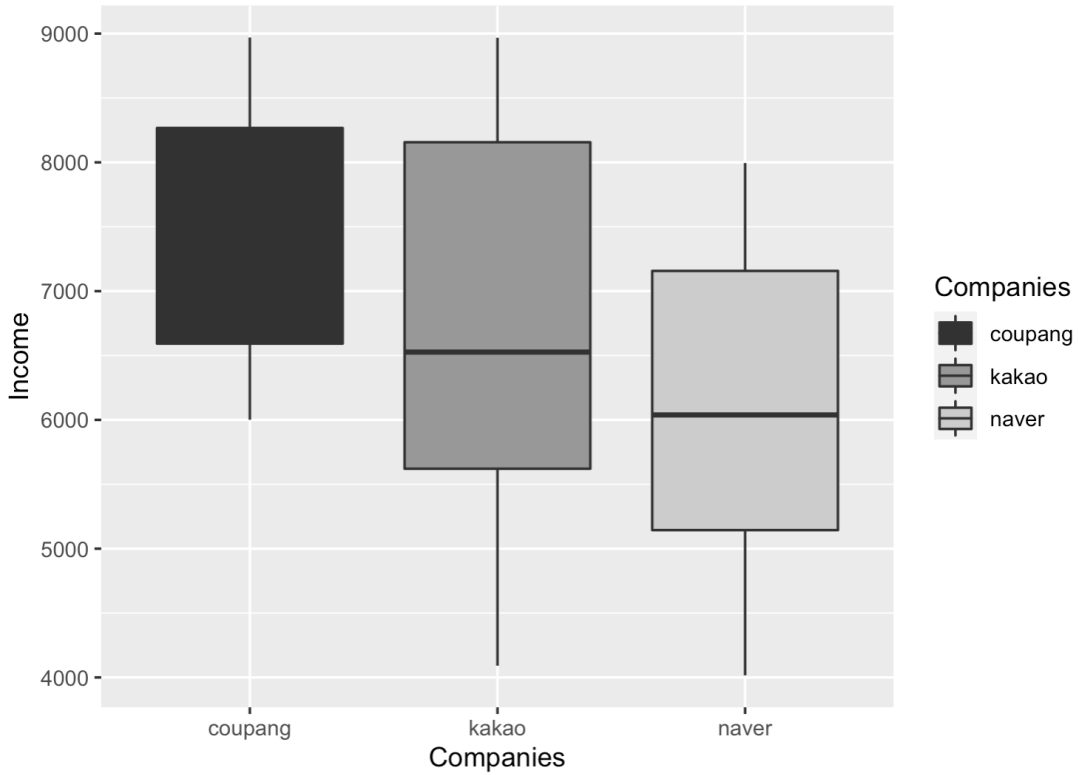
```
> library(dplyr)
> overall_2 <-  overall %>% mutate(Companies = ifelse(Company==1, "coupang", ifelse(Company==2, "kak
ao", "naver")))
> summary(overall_2)
    Company     Employee          Income      Companies
 Min.   :1   Min.   :  1.00   Min.   :4016   Length:300
 1st Qu.:1   1st Qu.: 25.75   1st Qu.:5774   Class :character
 Median :2   Median : 50.50   Median :6777   Mode  :character
 Mean   :2   Mean   : 50.50   Mean   :6738
 3rd Qu.:3   3rd Qu.: 75.25   3rd Qu.:7811
 Max.   :3   Max.   :100.00   Max.   :8970
```

문자 데이터를 사용한 상자 그림




```
> #문자 데이터를 사용한 상자 그림; 회사 이름을 사용해서 다시 상자그림 그리기
> ggplot(overall_2, aes(x=Companies,y=Income)) + geom_boxplot()
```

색상 지정하기




```
> # 상자 그림 회색 계열로 색상지정 (논문)
> ggplot(overall_2, aes(x=Companies,y=Income, fill=Companies)) + geom_boxplot() + scale_fill_grey()
```