기말과제_201821479_황혜린

201821479_황혜린

2019 5 12

목차

---

0. 가설

연습생들 순위에 따라 프로듀스X101 공식 트위터 계정에서의 연습생 별 언급횟수는 다를 것이다

0-1. 가설 수립 배경

> 요즘 프로듀스101의 새로운 시즌이 시작하면서 많은 화제를 불러모으고 있다.
> 본인도 바쁜 와중에 틈틈이 영상을 찾아보며 이번엔 누가 데뷔를 하게 될지 기대하고 있다.
> 프로듀스X101 공식 트위터를 보던 중 문득 연습생들 순위에 따라 연습생을 언급하는 횟수가 얼마나 차이나고, 만약 차이가 난다면 그 얼마나 영향을 미치는지 궁금해졌다. 공식계정은 특정 연습생을 홍보하는 용도가 아닌 방송 내용과 연습생들의 프로필, 소식, 연습생들의 간단한 동영상이 올라오기 때문이다. 따라서 8화 순위식까지의 프로듀스X101 공식계정을 크롤링해 가설을 검증해 보았다.

---

1. 데이터

1-1. 데이터 구하기 - 트위터 크롤링

```
#install.packages("twitteR")
library(twitteR)
library(ggplot2)
library(RColorBrewer)
library(tm)
```

```
## Loading required package: NLP
```

```
##
## Attaching package: 'NLP'
```

```
## The following object is masked from 'package:ggplot2':
##
##     annotate
```

```
library(syuzhet)

CONSUMER_SECRET <- "QrvUamEiqsPTVo9bXhSIcgUNWvDQyJdq2h9wj1zB4StPEkaKB1"
CONSUMER_KEY <- "d6ZZZQ7OKjDN9mq9x7Crkvh8W"
ACCESS_SECRET <- "xcVnyG0ejY3bykPvUnJfexoccEd3rpEaeSmRao8co3bvr"
ACCESS_TOKEN <- "1122681466649862152-66720Say10XNvfNbVyJU1asomSzySF"
setup_twitter_oauth(consumer_key=CONSUMER_KEY, consumer_secret = CONSUMER_SECRET,
                    access_token = ACCESS_TOKEN, access_secret = ACCESS_SECRET)
```

```
## [1] "Using direct authentication"
```

```
#프로듀스101 공식 트위터 계정 트윗 모두 크롤링
twitterUser <- getUser("mnet_produce101")
tweet.prdc <- userTimeline(twitterUser, n=2000)
tweet.prdc[[1]]$getClass()
```

```
## Reference Class "status":
##
## Class fields:
##
## Name:           text     favorited favoriteCount     replyToSN
## Class:     character       logical       numeric     character
##
## Name:        created     truncated    replyToSID            id
## Class:       POSIXct       logical     character     character
##
## Name:     replyToUID  statusSource    screenName  retweetCount
## Class:     character     character     character       numeric
##
## Name:      isRetweet     retweeted     longitude      latitude
## Class:       logical       logical     character     character
##
## Name:           urls
## Class:    data.frame
##
## Class Methods:
##      "setUrls", "getRetweets", "getRefClass", "getUrls", "setTruncated",
##      "setText", "getReplyToSID", "getText", "export", "setCreated",
##      "setFavoriteCount", "getCreated", "initialize", "callSuper",
##      "getRetweeters", "initFields", "getClass", "setReplyToUID", "import",
##      "setLatitude", "setIsRetweet", "getFavoriteCount", "getRetweetCount",
##      "getIsRetweet", "setId", "setScreenName", "getLatitude",
##      "getScreenName", "toDataFrame#twitterObj", "setRetweetCount",
##      "setReplyToSID", "getId", "getReplyToUID", "setFavorited",
##      "getRetweeted", "getFavorited", "toDataFrame", "setStatusSource",
##      "setReplyToSN", "copy", "usingMethods", "setRetweeted", "field",
##      ".objectParent", "getTruncated", "untrace", "trace", "setLongitude",
##      "getLongitude", "getStatusSource", ".objectPackage", "getReplyToSN",
##      "show"
##
## Reference Superclasses:
##      "twitterObj", "envRefClass"
```

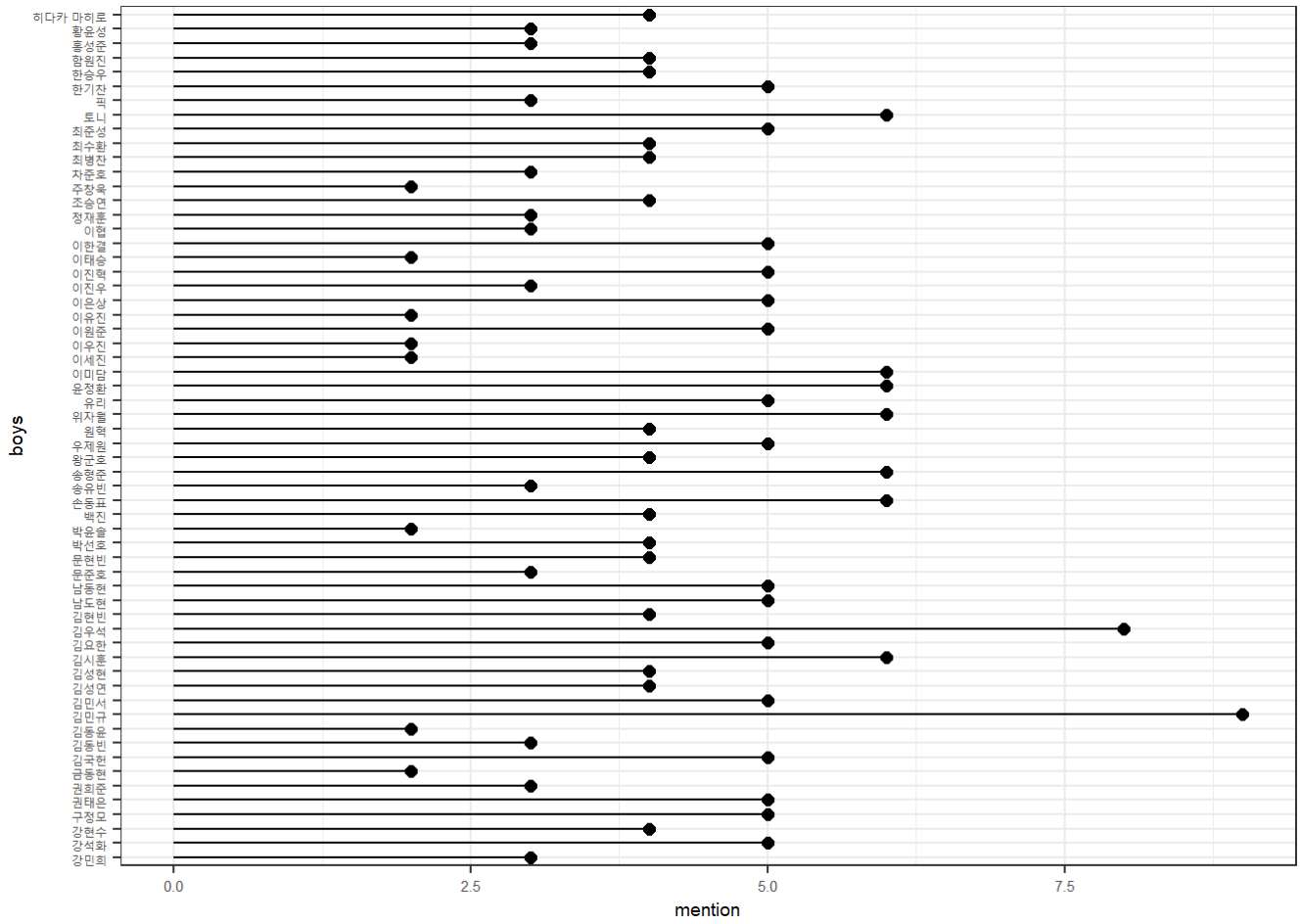
```
tweet.prdc <- twListToDF(tweet.prdc)


#프로듀스X101 트윗만 분류(다른 시즌을 제외 + 방영 첫날부터 순위발표식이 있던 날짜까지)
prdcX101 <- tweet.prdc[16:610,]
```

1-2. 데이터 정제

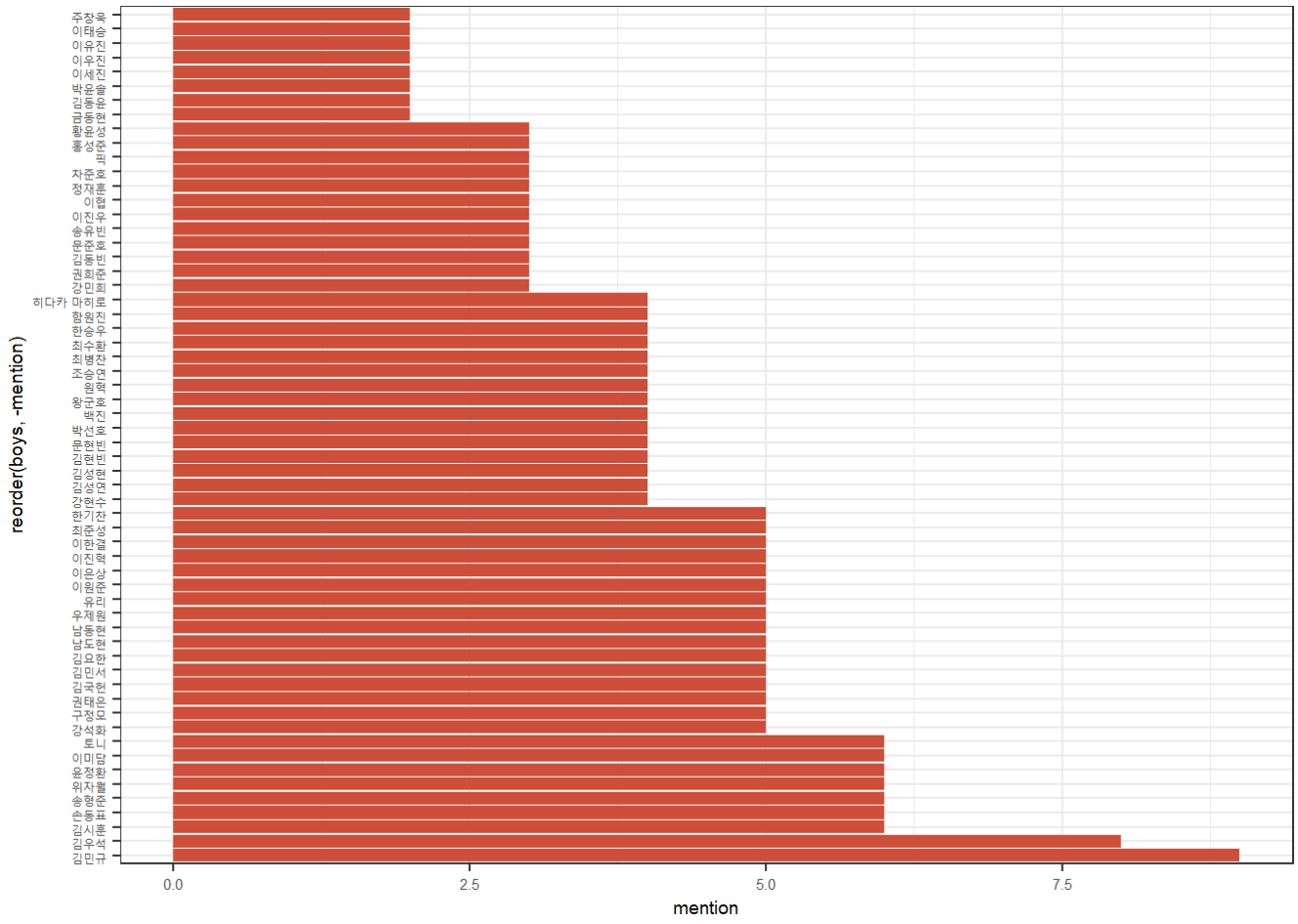
```
x <- c("김우석", "송형준", "김민규", "이진우", "김요한", "이은상", "남도현", "구정모", "송유빈", "함원진", "이진혁", "손동표", "한승우", "최병찬", "차준호", "금동현", "이한결", "김국헌", "조승연", "이세진", "황윤성", "강현수", "강민희", "토니", "김시훈", "박선호", "김현빈", "최수환", "김동윤", "백진", "주창욱", "권태은", "유리", "강석화", "윤정환", "이원준", "김동빈", "이미담", "문현빈", "이협", "최준성", "이우진", "박윤솔", "김성현", "위자월", "한기찬", "정재훈", "원혁", "김성연", "우제원", "픽", "권희준", "이유진", "김민서", "이태승", "홍성준", "남동현", "히다카 마히로", "문준호", "왕군호")

#데이터 프레임 생성
prdcX101.anov <- data.frame(matrix(nrow = 60, ncol = 0))
rownames(prdcX101.anov) <- x

#sapply에 적용시키기 위해 연습생들 이름으로 이뤄진 열 생성
prdcX101.anov$boys <- x

#연습생 별 언급횟수 구하기
prdcX101.anov$mention <- sapply(prdcX101.anov$boys, function(x){
  b <- grep(x, prdcX101$text)
  length(b)
})

#픽을 포함하는 단어의 횟수를 추출할 때, 연습생 "픽"만 나오는 것이 아닌 "원픽"과 "사심픽"도 나오므로 두 단어의 횟수를 제거하여 연습생의 언급횟수만 남긴다
prod1 <- function(x){
  b <- grep(x, prdcX101$text)
  length(b)
}
p <- prod1("픽") - prod1("원픽") - prod1("사심픽")
prdcX101.anov[51,2] <- p

#실제 순위
prdcX101.anov$rank <- c(1,4,10,8,3,6,7,5,12,14,2,11,9,16,13,19,15,22,17,21,18,20,23,27,24,28,25,29,31,36,30,34,40,35,38,47,42,37,32,26,46,41,48,45,43,57,39,33,44,54,50,56,55,52,53,51,60,49,59,58)
```

1-3. 데이터 시각화

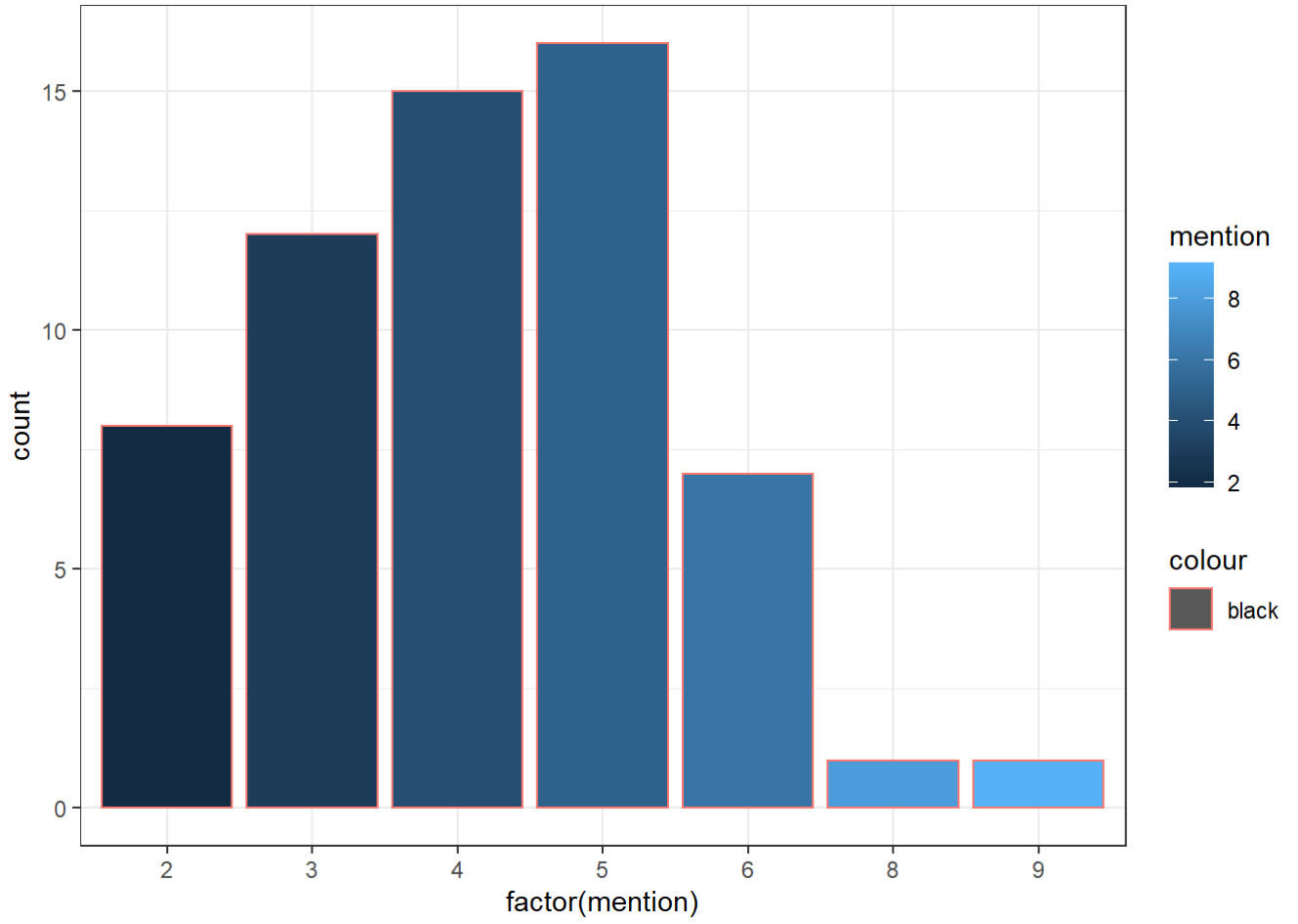
```
#연습생들 별 언급 횟수
library(ggplot2)
theme_set(theme_bw())
ggplot(data = prdcX101.anov,aes(x=boys, y=mention)) + geom_point(size=2)+ geom_segment(aes(x=boys, xend=boys, y=0, yend=mention)) + theme(text = element_text(size=7)) + coord_flip()
```




```
#연습생들 별 언급횟수(내림차순)
ggplot(data = prdcX101.anov, aes(x=reorder(boys,-mention), y=mention)) + geom_bar(stat = "identity", f
ill="tomato3") + theme(text = element_text(size=7)) + coord_flip()
```




```
#언급 횟수별 연습생 수
ggplot(data = prdcX101.anov, aes(factor(mention))) + geom_bar(aes(fill=mention, colour="black"))
```



2. 통계분석

```
#분석 준비
#연습생들을 순위별로 10등씩 6개 집단으로 분류
prdcX101.anov$rank.grade <- sapply(prdcX101.anov$rank, function(x){
  ((x-1)%/%10)+1
})

#독립변수를 범주화
prdcX101.anov$rank.grade.factor <- factor(prdcX101.anov$rank.grade, levels = c(1,2,3,4,5,6), labels =
 c("1등-10등", "11등-20등", "21등-30등", "31등-40등", "41등-50등", "51등-60등"))
str(prdcX101.anov$rank.grade.factor)
```

```
##  Factor w/ 6 levels "1등-10등","11등-20등",..: 1 1 1 1 1 1 1 1 2 2 ...
```

2-1. 기술통계

```
#빈도수
table(prdcX101.anov$mention)
```

```
##
##  2  3  4  5  6  8  9
##  8 12 15 16  7  1  1
```

```
#5 number summary
summary(prdcX101.anov)
```

```
##      boys              mention          rank         rank.grade
##  Length:60          Min.   :2.000   Min.   : 1.00   Min.   :1.0
##  Class :character   1st Qu.:3.000   1st Qu.:15.75   1st Qu.:2.0
##  Mode  :character   Median :4.000   Median :30.50   Median :3.5
##                     Mean   :4.183   Mean   :30.50   Mean   :3.5
##                     3rd Qu.:5.000   3rd Qu.:45.25   3rd Qu.:5.0
##                     Max.   :9.000   Max.   :60.00   Max.   :6.0
##  rank.grade.factor
##  1등-10등 :10
##  11등-20등:10
##  21등-30등:10
##  31등-40등:10
##  41등-50등:10
##  51등-60등:10
```

```
#빈도표
cbind(`빈도` = table(prdcX101.anov$mention),
       `누적빈도` = cumsum(table(prdcX101.anov$mention)),
         `비율` = prop.table(table(prdcX101.anov$mention)),
  `누적비율` = cumsum(prop.table(table(prdcX101.anov$mention))))
```

```
##   빈도 누적빈도       비율  누적비율
## 2    8        8 0.13333333 0.1333333
## 3   12       20 0.20000000 0.3333333
## 4   15       35 0.25000000 0.5833333
## 5   16       51 0.26666667 0.8500000
## 6    7       58 0.11666667 0.9666667
## 8    1       59 0.01666667 0.9833333
## 9    1       60 0.01666667 1.0000000
```

2-2. 추론통계 - 분산분석

```
#분산분석 실행
anov <- aov(mention ~ rank.grade.factor, data = prdcX101.anov)
anova(anov)
```

```
## Analysis of Variance Table
##
## Response: mention
##                  Df  Sum Sq Mean Sq F value  Pr(>F)
## rank.grade.factor  5  23.883  4.7767  2.5018 0.04142 *
## Residuals         54 103.100  1.9093
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- 집단 간 제곱합은 23.883이고, 자유도는 5이므로 집단 간 평균 제곱합은 4.7767
- 집단 내 제곱합은 103.100이고, 자유도는 54이므로 집단 내 평균 제곱합은 1.9093
- 유의도는 0.04142로 0.05보다 작으므로 영가설을 기각한다.

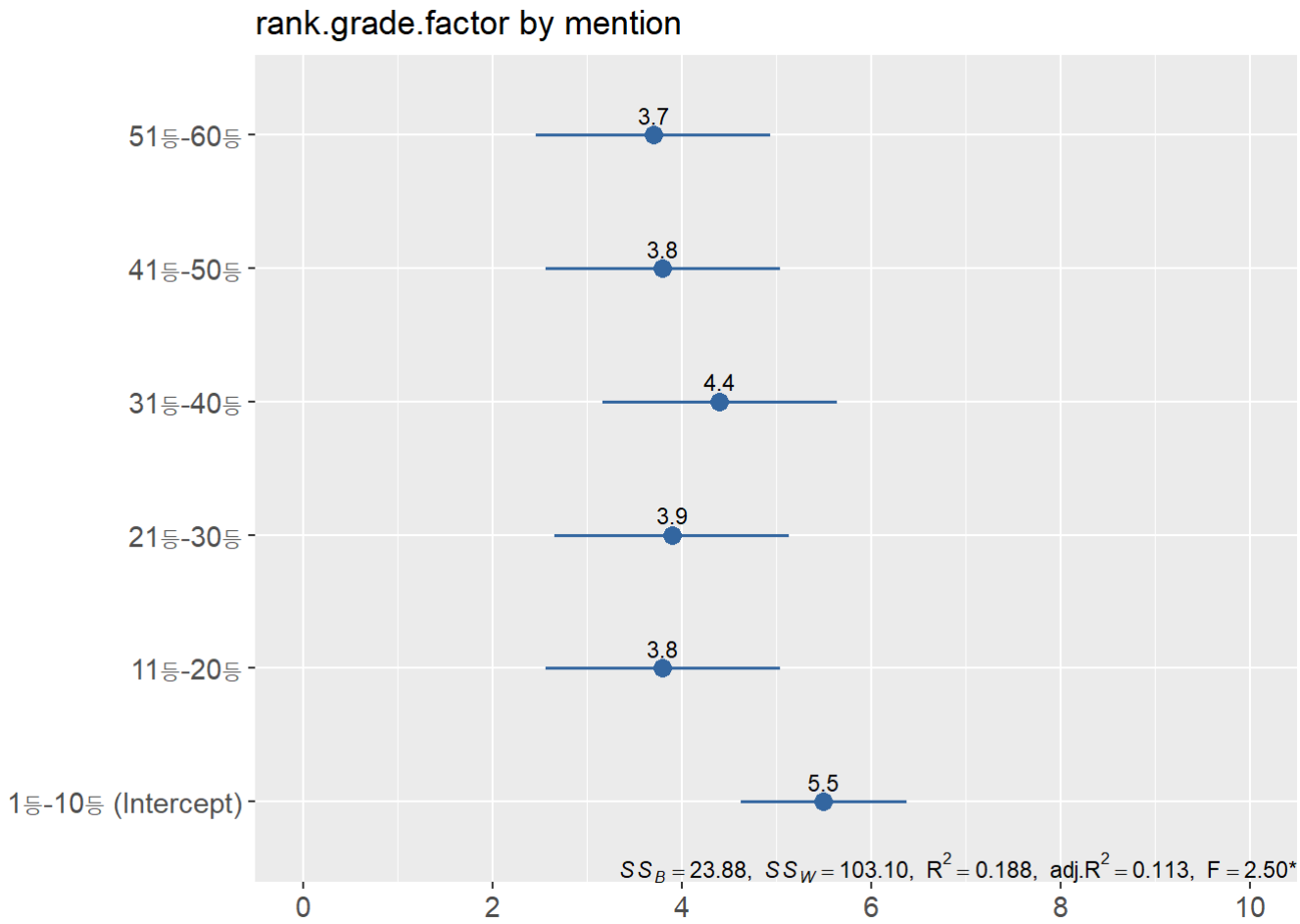
따라서 “연습생들 순위에 따라 프로듀스X101 공식 트위터 계정에서의 연습생 별 언급 횟수는 다를 것이다”는 연구가설을 채택한다.

2-3. sjPlot패키지를 이용한 분산분석 그래프

```
library(sjPlot)
```

```
## Install package "strengejacke" from GitHub (`devtools::install_github("strengejacke/strengejacke")
`) to load all sj-packages at once!
```

```
set_theme(axis.textsize =1, geom.label.size =3)
sjp.aov1(prdcX101.anov$mention, prdcX101.anov$rank.grade.factor, meansums = T,
show.summary = T, show.p =F, geom.size
= 3)
```



- 각 집단에 표시된 선의 길이는 신뢰구간을 의미한다.
- 집단의 순위의 평균이 표시되어 출력된다.

2-4. 사후검증

```
#사후검증
library(agricolae)
scheffe.test(anov, "rank.grade.factor", alpha = 0.05, console = T)
```

```
##
## Study: anov ~ "rank.grade.factor"
##
## Scheffe Test for mention
##
## Mean Square Error  : 1.909259
##
## rank.grade.factor,  means
##
##           mention      std  r Min Max
## 11등-20등     3.8 1.135292 10   2   6
## 1등-10등      5.5 1.779513 10   3   9
## 21등-30등     3.9 1.449138 10   2   6
## 31등-40등     4.4 1.264911 10   2   6
## 41등-50등     3.8 1.316561 10   2   6
## 51등-60등     3.7 1.251666 10   2   5
##
## Alpha: 0.05 ; DF Error: 54
## Critical Value of F: 2.38607
##
## Minimum Significant Difference: 2.134391
##
## Means with the same letter are not significantly different.
##
##           mention groups
## 1등-10등      5.5      a
## 31등-40등     4.4      a
## 21등-30등     3.9      a
## 11등-20등     3.8      a
## 41등-50등     3.8      a
## 51등-60등     3.7      a
```

- 사후검증 결과, 집단은 a로만 나누어졌다. 즉, 독립변수의 6집단의 평균이 유의한 차이를 보이지 않는다.

3. 제언

연구가설을 채택하는 결과가 나왔다. 예상은 했지만 연습생들의 순위에 따라 언급 횟수가 다르다는 사실이 어쩔 수 없나 싶으면서도 씁쓸하다.
보통 예측만 하고 지나가는 경우가 대부분인데, 직접 데이터를 구하여 프로그래밍 언어로 지표를 시각화하고 통계기법을 활용한 분석으로 내 의견을 뒷받침해줄 수 있다는 것이 데이터 분석의 장점인 것 같다.앞으로도 해결할 문제가 있다면 데이터 분석을 통해 문제를 해결해보고 싶다.