패키지 오답노트 2회차

서희나

2023-01-21

Part1. Preprocessing

문제 0. (기본 세팅) 아래의 코드를 시행해 주세요.

- library(tidyverse)
- library(data.table)
- library(magrittr)
- data = fread(“All Categories.csv”)

```
pacman::p_load(tidyverse,
               data.table,
               magrittr)
```

```
data = fread("All Categories.csv")
```

문제 1. 데이터의 구조를 자유롭게 파악하세요

```
data %>% glimpse
```

```
## Rows: 600
## Columns: 9
## $ Rank          <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 1…
## $ Name          <chr> "Meesho: Online Shopping App", "Shopee: Online Shopping"…
## $ Developer     <chr> "Meesho", "Shopee", "Instagram", "MX Media (formerly J2 …
## $ Category      <chr> "Shopping", "Shopping", "Social", "Video Players & Edito…
## $ Size          <chr> "15 MB", "68 MB", "41 MB", "36 MB", "12 MB", "64 MB", "4…
## $ `Star Rating` <dbl> 4.4, 4.1, 4.3, 4.1, 4.5, 4.2, 4.0, 4.3, 4.3, 3.8, 4.3, 4…
## $ Reviews       <chr> "15L", "76T", "13Cr", "1Cr", "41T", "2Cr", "34L", "2Cr",…
## $ Downloads     <chr> "10Cr+", "1Cr+", "100Cr+", "100Cr+", "1Cr+", "100Cr+", "…
## $ `Rated for`   <chr> "3+", "3+", "12+", "3+", "3+", "12+", "3+", "12+", "12+"…
```

```
data %>%
  select_if(summarise_all(.,n_distinct)<100) %>%
  apply(2,unique)
```

```
## $Category
##  [1] "Shopping"                "Social"
##  [3] "Video Players & Editors" "Tools"
##  [5] "Communication"           "Business"
##  [7] "Finance"                 "Productivity"
##  [9] "Music & Audio"           "Entertainment"
## [11] "News & Magazines"        "Photography"
## [13] "Travel & Local"          "Food & Drink"
## [15] "Auto & Vehicles"         "Education"
## [17] "Books & Reference"       "Art & Design"
## [19] "Sports"                  "Personalisation"
## [21] "Lifestyle"               "Health & Fitness"
## [23] "Maps & Navigation"       "Dating"
## [25] "Medical"                 "Comics"
## [27] "Weather"                 "House & Home"
##
## $`Star Rating`
##  [1] "4.4" "4.1" "4.3" "4.5" "4.2" "4.0" "3.8" "4.6" "3.5" "3.0" "3.9" "3.7"
## [13] "3.6" "4.7" "3.4" "3.1" "2.1" "3.3" "4.8" NA    "2.8" "2.6" "3.2" "4.9"
##
## $Downloads
##  [1] "10Cr+"  "1Cr+"   "100Cr+" "50Cr+"  "5Cr+"   "500Cr+" "50L+"   "10L+"
##  [9] "1L+"    "5L+"    "1TCr+"  "50T+"   "10T+"
##
## $`Rated for`
## [1] "3+"  "12+" "18+" "16+" "7+"
```

문제 2. 각 변수 별 NA 의 개수를 확인하고 NA 가 존재하는 행을 제거합니다.

```
data %>% is.na %>% colSums
```

```
##        Rank        Name   Developer    Category        Size Star Rating
##           0           0           0           0           0           1
##     Reviews   Downloads   Rated for
##           0           0           0
```

```
data %>% is.na %>% sum
```

```
## [1] 1
```

```
data %<>% na.omit
sum(is.na(data))
```

```
## [1] 0
```

문제 3. 각 변수 별 unique 의 개수를 확인하세요.

```
data %>% summarise_all(n_distinct)
```

```
##   Rank Name Developer Category Size Star Rating Reviews Downloads Rated for
## 1  599  599       491       28  137          23     166        13         5
```

문제 4. 문자형 변수 중 범주형인 변수를 범주형으로 변경해주세요.

```
data %<>% mutate_at(vars(Downloads,Category,`Rated for`),as.factor)
```

1. 범주형 변수만 unique 값을 출력해 이상한 값이 없는지 먼저 확인해봅니다. Downloads 가 “1TCr+”인 행은 제거하겠습니다

```
data %>% select_if(is.factor) %>% lapply(unique)
```

```
## $Category
##  [1] Shopping               Social                 Video Players & Editors
##  [4] Tools                  Communication          Business
##  [7] Finance                Productivity           Music & Audio
## [10] Entertainment          News & Magazines       Photography
## [13] Travel & Local         Food & Drink           Auto & Vehicles
## [16] Education              Books & Reference      Art & Design
## [19] Sports                 Personalisation        Lifestyle
## [22] Health & Fitness       Maps & Navigation      Dating
## [25] Medical                Comics                 Weather
## [28] House & Home
## 28 Levels: Art & Design Auto & Vehicles Books & Reference Business ... Weather
##
## $Downloads
##  [1] 10Cr+  1Cr+   100Cr+ 50Cr+  5Cr+   500Cr+ 50L+   10L+   1L+    5L+
## [11] 1TCr+  50T+   10T+
## 13 Levels: 100Cr+ 10Cr+ 10L+ 10T+ 1Cr+ 1L+ 1TCr+ 500Cr+ 50Cr+ 50L+ ... 5L+
##
## $`Rated for`
## [1] 3+  12+ 18+ 16+ 7+
## Levels: 12+ 16+ 18+ 3+ 7+
```

```
data %<>% filter(!data$Downloads=='1TCr+')
sum(data$Downloads=='1TCr+')
```

```
## [1] 0
```

3. 이때, Rated_for 와 Downloads 는 level 의 순서를 지정해주겠습니다.

- Rated_for à “3+”,“7+”,“12+”,“16+”,“18+”
- Downloadsà”10T+“,”50T+“,”1L+“,”5L+“,”10L+“,”50L+“,”1Cr+“,”5Cr+“,”10Cr+“,”50Cr+“,”100Cr+“,”500Cr+”

```
# Rated_for -> "3+","7+","12+","16+","18+"
data$`Rated for` %<>% factor(levels=c("3+","7+","12+","16+","18+"),order= T)

# Downloads

data$Downloads %<>% factor(levels=c("10T+", "50T+", "1L+", "5L+", "10L+", "50L+", "1Cr+", "5Cr
+", "10Cr+","50Cr+", "100Cr+", "500Cr+"),order=T)
```

```
str(data$`Rated for`)
```

```
##  Ord.factor w/ 5 levels "3+"<"7+"<"12+"<..: 1 1 3 1 1 3 1 3 3 3 ...
```

```
str(data$Downloads)
```

```
##  Ord.factor w/ 12 levels "10T+"<"50T+"<..: 9 7 11 11 7 11 10 9 11 8 ...
```

문제 5. 단위로 표현된 Review 변수를 각 단위에 맞는 수를 곱해주어 수치형 변수를 만들어주세요.

이때, 새롭게 생성한 변수명은 review_num 입니다.

1. 먼저, 데이터 설명서에서 제시된 단위 이외의 단위가 존재하는지 확인합니다.(str_extract)

```
data$Reviews %>%str_extract(.,'[A-z]+') %>% unique
```

```
## [1] "L"  "T"  "Cr" NA
```

2. 단위에 따라서 숫자를 바꿔줄 때 각 단위에 따라 적절한 숫자를 곱해주는 review_to_num 함수를 직접 만드세요

```
review_to_num<-function(x){
  options(scipen=999)
  if(is.na(str_extract(x,'[A-z]+'))==T){num<-as.numeric(x)}
  else if(str_extract(x,'[A-z]+')=='L'){
    num<-as.numeric(gsub('WWD','', x))*100000}
  else if(str_extract(x,'[A-z]+')=='T'){
    num<-as.numeric(gsub('WWD','', x))*1000}
  else{num<-as.numeric(gsub('WWD','', x))*10000000}

  return(num)
}
```

```
review_to_num('46Cr')
```

```
## [1] 460000000
```

3. review_to_num 함수를 이용해 data 에 review_num 열을 추가로 생성합니다.

```
data %<>% mutate(review_num=
                   sapply(data$Reviews,review_to_num) %>%
                   as.vector)
```

문제 6. Size 도 review 와 비슷한 과정을 거쳐 수정하겠습니다.

1. Size 의 단위에는 어떠한 것들이 있는지 확인해보세요.(str_extract)

```
data$Size %>% str_extract(.,'[A-z]+') %>% unique
```

```
## [1] "MB"      "Teacher" "KB"
```

2. 단위의 종류를 확인한 결과 용량이 아닌 행이 있습니다. 이 행을 출력해보세요.

```
data %>% filter(str_extract(data$Size,'[A-z]+')=="Teacher")
```

```
##    Rank         Name  Developer      Category             Size Star Rating
## 1:  153 YouTube Kids Google LLC Entertainment Teacher Approved           4
##    Reviews Downloads Rated for review_num
## 1:      18L     10Cr+        3+    1800000
```

```
data[which(str_extract(data$Size,'[A-z]+')=="Teacher"),]
```

```
##    Rank         Name  Developer      Category             Size Star Rating
## 1:  153 YouTube Kids Google LLC Entertainment Teacher Approved           4
##    Reviews Downloads Rated for review_num
## 1:      18L     10Cr+        3+    1800000
```

이전의 NA 가 존재하는 행이었습니다. 여기서 이 행을 제거하겠습니다.

```
data %<>% filter(!str_extract(data$Size,'[A-z]+')=="Teacher")
```

3. 1MB = 1024KB 입니다. KB 단위인 행이 더 많은지, MB 인 행이 더 많은지 확인해보고 더 많이 존재하는 단위로 Size 를 통일 시킨 후 다시 단위를 제거하고 수치형 변수로 만들어주세요. (사이즈 통일과 수치형 변환의 순서는 상관 없습니다!)

```
data$Size %>% str_extract(.,'[A-z]+') %>% table
```

```
## .
##  KB  MB
##   2 595
```

```
unit_to_num<-function(x){
  options(scipen=999)
  if(str_extract(x,'[A-z]+')=='MB'){
    num<-as.numeric(gsub('MB','', x))}
  else{num<-as.numeric(gsub('MB','', x))*1/1024}

  return(num)
}
```

```
data %<>% mutate_at(vars(Size),
                    function(x){sapply(x,unit_to_num) %>% as.vector})
```

문제 7. 다음 파트에서는 시각화 연습을 할텐데요, 그 전에 상위 600 개에 속한 어플 중 가장 많은 Category 6 개에 속하는 어플만 있는 데이터프레임을 생성해주고, top6_df 라는 이름으로 저장해주세요.

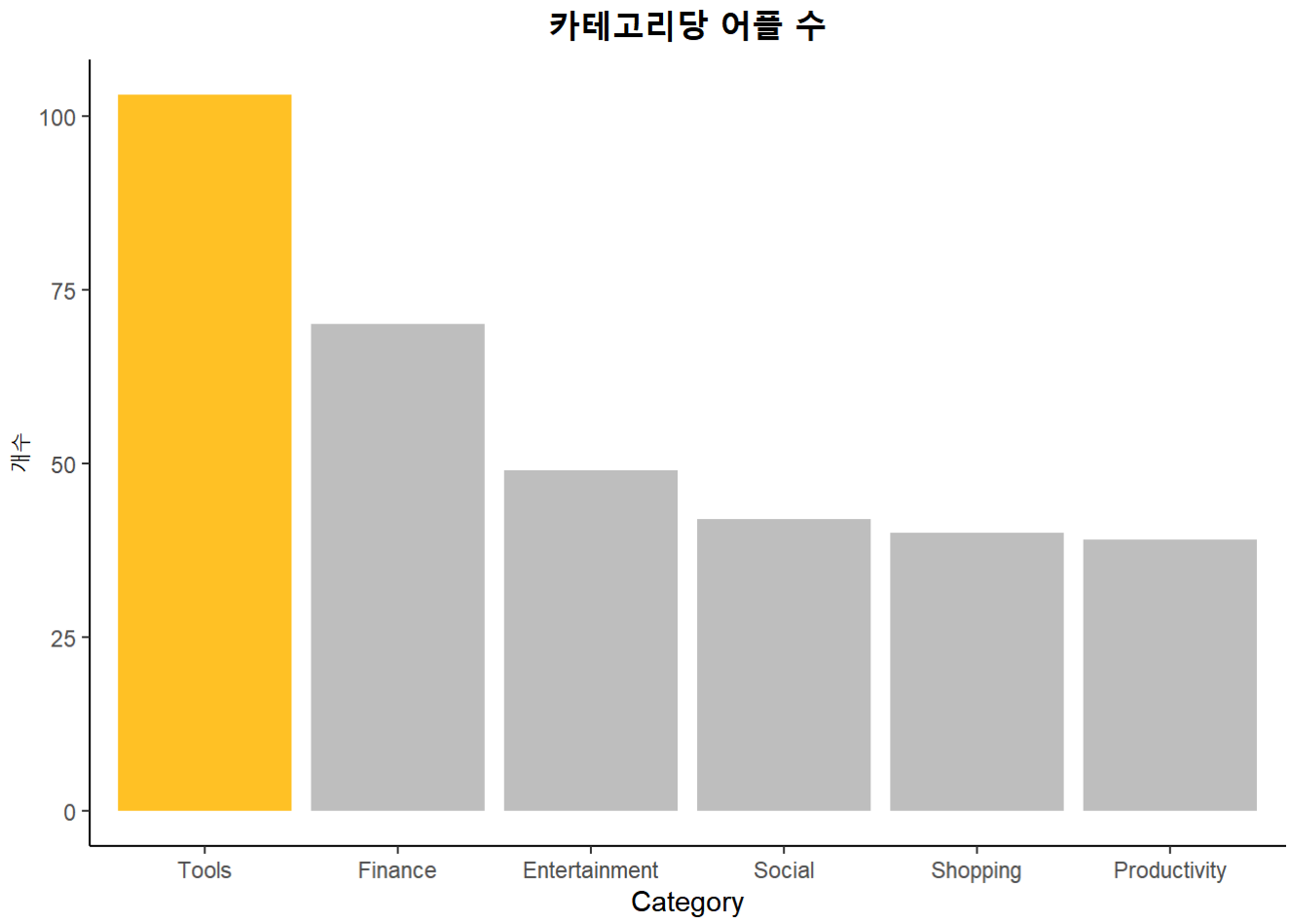
```
top6<-data %>% group_by(Category) %>%
  summarise(num=n()) %>%
  arrange(-num) %>%
  mutate(index=c(1:length(Category))) %>%
  filter(index<=6) %>%
  select(Category) %>% unlist

top6_df<-data %>% filter(data$Category %in% top6)
```

Part 2. Visualization

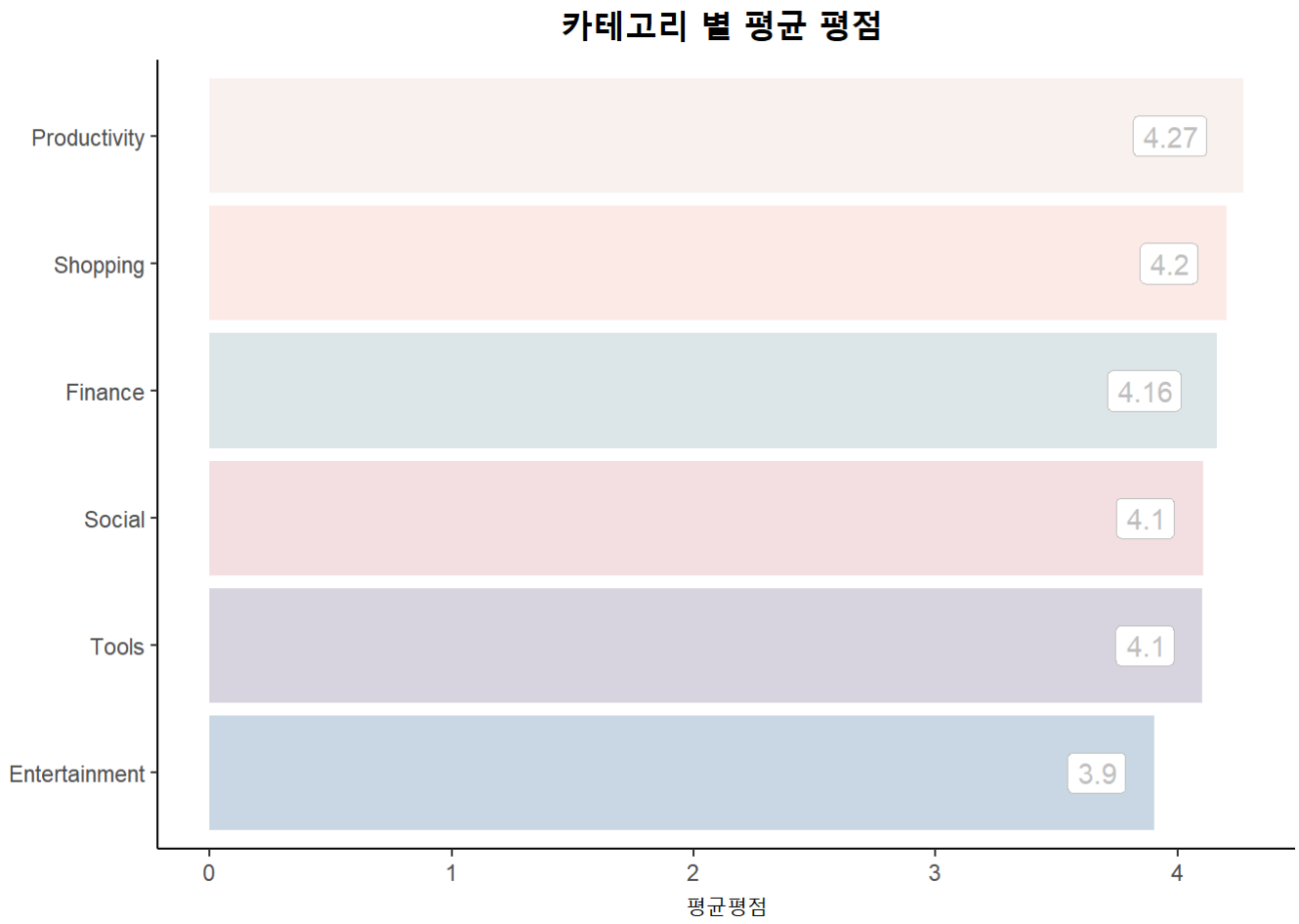
문제 1. 아래의 조건에 맞춰 다음과 같은 그래프를 완성해주세요

```
top6_df %>%
  group_by(Category) %>%
  summarise(num=n()) %>%
  ggplot(aes(x=reorder(Category,-num),y=num,fill=reorder(Category,-num)))+
  geom_bar(stat='identity')+
  labs(title='카테고리당 어플 수',
       x='Category',
       y='개수')+
  scale_fill_manual(values=c('goldenrod1',rep('grey',5)))+
  theme_classic()+
  theme(plot.title=element_text(hjust=0.5,face='bold',size=15),
        legend.position='none')
```



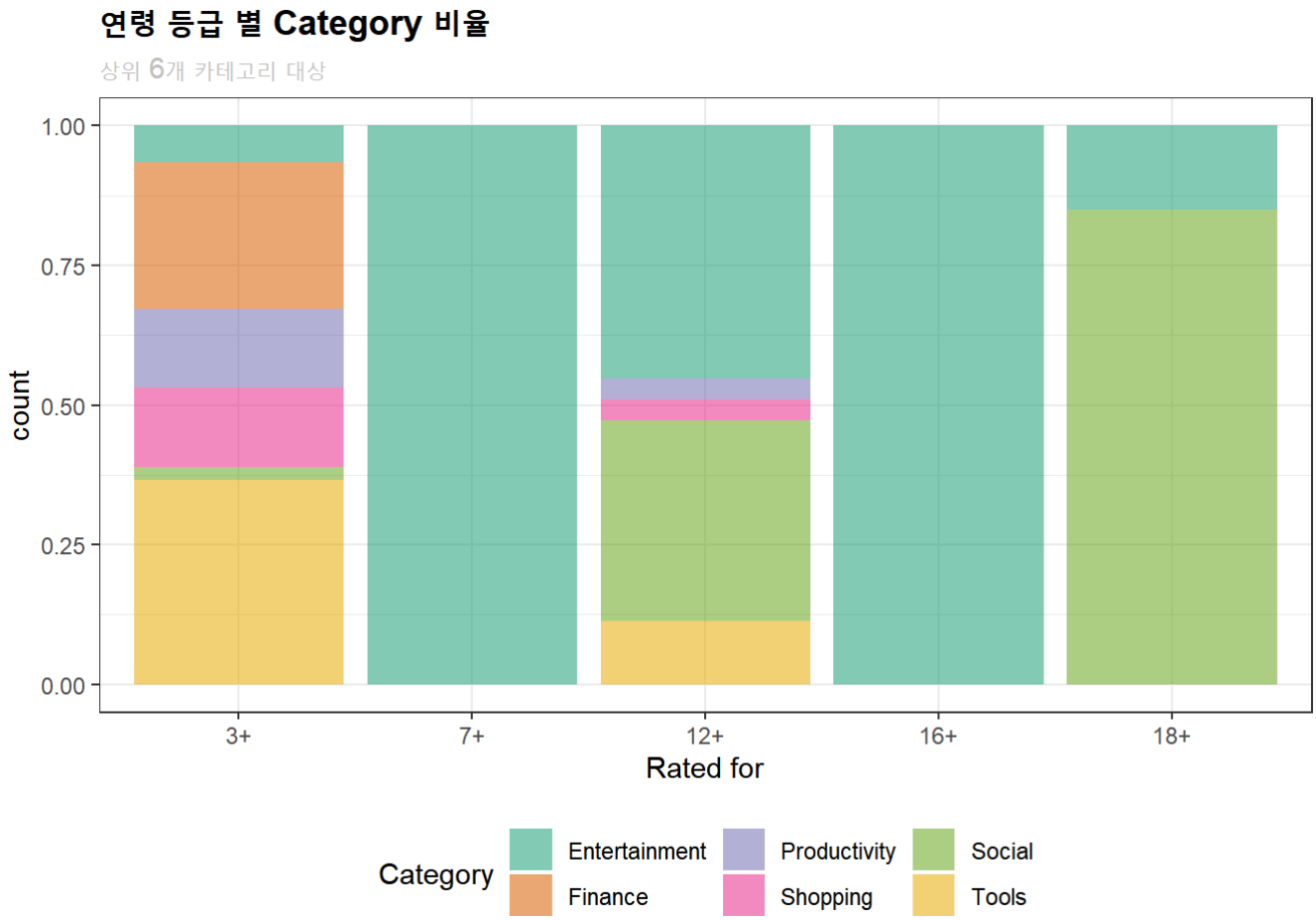
문제 2. 아래의 조건에 맞춰 다음과 같은 그래프를 완성해주세요.

```
top6_df %>%
  group_by(Category) %>%
  summarise(average=mean(`Star Rating`)) %>%
  ggplot(aes(x=reorder(Category,average),y=average,
             fill=Category))+
  geom_bar(stat='identity',alpha=0.3)+
  scale_fill_manual(values=c("#4878A1","#8AA8AF", "#E7CFC5","#F4BAAB","#D8959E","#7B7092"))+
  theme_classic()+
  coord_flip()+
  labs(title='카테고리 별 평균 평점 ',
       y='평균평점',
       x=NULL)+
  theme(plot.title=element_text(size=15,face='bold',hjust=0.5),
        axis.title.y=element_text(size=10),
        legend.position='none')+
  geom_label(aes(label=average %>% round(2)),
             color='grey',fill='white',hjust=1.5)
```



문제 3.

```
top6_df %>%
  ggplot(aes(x=`Rated for`,fill=Category))+
  geom_bar(position='fill',alpha=0.55)+
  scale_fill_brewer(palette='Dark2')+
  theme_bw()+
  labs(title='연령 등급 별 Category 비율',
       subtitle='상위 6개 카테고리 대상')+
  theme(plot.title=element_text(face='bold'),
        plot.subtitle=element_text(color='grey'),
        legend.position='bottom')
```

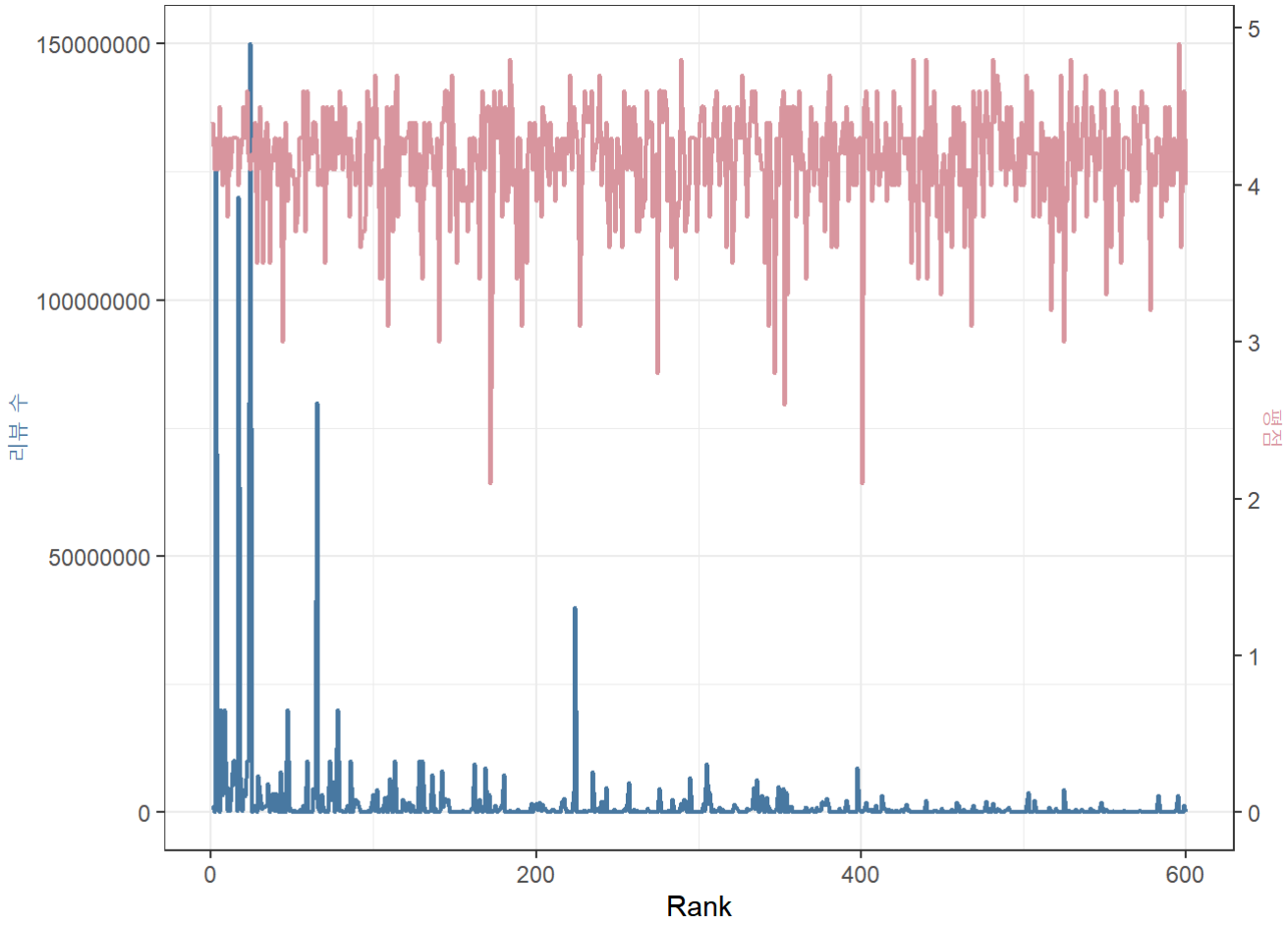


문제 4.

```
ratio<-max(data$review_num)/max(data$`Star Rating`)

data %>% ggplot(aes(x=Rank))+
  geom_line(aes(y=review_num),color='#4878A1',size=0.71)+
  geom_line(aes(y=ratio*`Star Rating`),color='#D8959E',size=0.71)+
  scale_y_continuous(name='리뷰 수',
                     sec.axis=sec_axis(trans=~./ratio,name='평점'))+
  theme_bw()+
  theme(axis.title.y.left = element_text(color='#4878A1'),
        axis.title.y.right = element_text(color='#D8959E'))
```

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## ℹ Please use `linewidth` instead.
```


150000000
100000000
50000000
0
리뷰 수
5
4
3
2
1
0
평점
0
200
400
600
Rank