

TEST_02

- Guide
 - 1. Data Wrangling Problem
 - 2. Visualization Problem(Using ggplot2)
-

- 필요 패키지 : `dplyr`, `tidyr`, `ggplot2`
 - 필요 데이터 : `Sales.csv`
-

Guide

- 시험시간 : 13:30 ~ 15:00 (총90분)
- R markdown을 이용하여 생성한 html파일 모두 제출
- Dropbox에 제출시 html 파일과 rmd 파일 모두 업로드 해야함
 - 공유된 폴더안에 시험날짜 폴더를 만들고 (예 : 202101234)
 - 관련 데이터는 Data폴더 를 만들어 제출
 - html,rmd파일은 Rmd폴더 에 학번_이름으로 작성 제출 (예 : 202101234_홍길동)
- 문제에서 주어진 결과와 같게 작성
 - ex) 작성방식 : 1.1번 문제라고 적고 R Chunk를 생성하여 그 안에 해당 code 기입
- 업로드 및 수정시간이 모두 기록 됨에 유의
- 코드공유 적발시 0점 처리
- 문제에서 출력한대로 출력(전체데이터 출력은 지양)
- 위의 유의사항을 어길 시 감점
- 시험시간을 넘겨서 제출하면 0점 처리



0. Packages

```
library(dplyr)
library(tidyr)
library(ggplot2)
```

1. Data Wrangling Problem

1.1 Sales.csv 를 불러오시오.[5점]

- head()로 출력할 것

```
data <- read.csv(file = "../Data/Sales.csv")
head(data)
```

##	X	Order.ID	Product	Quantity.Ordered	Price.Each
## 1	1	141234	iPhone	1	700.00
## 2	2	141235	Lightning Charging Cable	1	14.95
## 3	3	141236	Wired Headphones	2	11.99
## 4	4	141237	27in FHD Monitor	1	149.99
## 5	5	141238	Wired Headphones	1	11.99
## 6	6	141239	AAA Batteries (4-pack)	1	2.99

##	Order.Date	Purchase.Address
## 1	01/22/19 21:25	944 Walnut St, Boston, MA 02215
## 2	01/28/19 14:15	185 Maple St, Portland, OR 97035
## 3	01/17/19 13:33	538 Adams St, San Francisco, CA 94016
## 4	01/05/19 20:33	738 10th St, Los Angeles, CA 90001
## 5	01/25/19 11:59	387 10th St, Austin, TX 73301
## 6	01/29/19 20:22	775 Willow St, San Francisco, CA 94016

1.2 데이터의 속성을 확인하시오[5점]

```
str(data)
```

```
## 'data.frame':    55041 obs. of  7 variables:
##  $ X                : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Order.ID         : int  141234 141235 141236 141237 141238 141239
141240 141241 141242 141243 ...
##  $ Product          : chr  "iPhone" "Lightning Charging Cable" "Wired
Headphones" "27in FHD Monitor" ...
##  $ Quantity.Ordered: int  1 1 2 1 1 1 1 1 1 1 ...
##  $ Price.Each       : num  700 14.9 12 150 12 ...
##  $ Order.Date       : chr  "01/22/19 21:25" "01/28/19 14:15" "01/17/1
9 13:33" "01/05/19 20:33" ...
##  $ Purchase.Address: chr  "944 Walnut St, Boston, MA 02215" "185 Map
le St, Portland, OR 97035" "538 Adams St, San Francisco, CA 94016" "738
10th St, Los Angeles, CA 90001" ...
```

1.3 다음과 같이 변수 값과 속성을 변경 후 str()을 통해 확인하시오.[25점]

- Order.Date 변수에서 월(month)만 추출하시오. [5점]
 - ex) 01/22/19 21:25에서 01만 추출
- Purchase.Address에서 가운데 이름만 가져오시오.[5점]
 - ex) 944 Walnut St, Boston, MA 02215에서 Boston만 추출
- 필요한 열만 가져오고 열 이름은 결과물과 같게 하시오. [5점]
- month의 속성을 변경하시오. [10점]
 - 힌트: factor를 활용할 것, 1월 ~ 4월을 각각 “Jan”, “Feb,”Mar“,”Apr”로 바꾸시오.

```
data=data %>% mutate(month = substr(Order.Date, 1, 2)) %>%
  separate(Purchase.Address, c("temp", "Address", "temp2"), sep=",") %>%
  select(c(Order.ID, Product, Quantity.Ordered, Price.Each, Address, month))
data$month = factor(data$month, levels=c("01","02","03","04"), labels =
c("Jan","Feb","Mar","Apr"))
str(data)
```

```
## 'data.frame':    55041 obs. of  6 variables:
##  $ Order.ID      : int  141234 141235 141236 141237 141238 141239
141240 141241 141242 141243 ...
##  $ Product       : chr  "iPhone" "Lightning Charging Cable" "Wired
Headphones" "27in FHD Monitor" ...
##  $ Quantity.Ordered: int   1 1 2 1 1 1 1 1 1 1 ...
##  $ Price.Each     : num   700 14.9 12 150 12 ...
##  $ Address        : chr   " Boston" " Portland" " San Francisco" " L
os Angeles" ...
##  $ month          : Factor w/ 4 levels "Jan","Feb","Mar",...: 1 1 1
1 1 1 1 1 1 ...
```

1.4 월별로 가장 많이 팔린 상위 3개 제품을 구하시오[15점]

- 힌트: Quantity.Ordered를 활용하여 count를 구하시오 [5점]
- 결과물 처럼 정렬하여 나타낼 것 [5점]
- 월별로 상위 3개 filtering [5점]

```
data %>% group_by(month, Product) %>% summarise(count=sum(Quantity.Orde
red)) %>% arrange(month, -count) %>% top_n(3)
```

```
## # A tibble: 12 x 3
## # Groups:   month [4]
##   month Product          count
##   <fct> <chr>          <int>
## 1 Jan   AAA Batteries (4-pack)  1595
## 2 Jan   AA Batteries (4-pack)   1424
## 3 Jan   USB-C Charging Cable    1283
## 4 Feb   AAA Batteries (4-pack)  1971
## 5 Feb   AA Batteries (4-pack)   1734
## 6 Feb   USB-C Charging Cable    1653
## 7 Mar   AAA Batteries (4-pack)  2475
## 8 Mar   AA Batteries (4-pack)   2214
## 9 Mar   USB-C Charging Cable    1943
## 10 Apr  AAA Batteries (4-pack)   2936
## 11 Apr  AA Batteries (4-pack)    2819
## 12 Apr  Lightning Charging Cable 2371
```

1.5 지역별 평균 소비액을 구하시오[15점]

- Quantity.Ordered와 Price.Each를 곱하여 “consume”이라는 열 생성 [5점]
- 지역별 평균 소비액을 구하시오 [5점]
- 정렬하여 나타내시오 [5점]

```
data %>% mutate(consume = Quantity.Ordered*Price.Each) %>% group_by(Address) %>% summarise(consume = mean(consume)) %>% arrange(-consume)
```

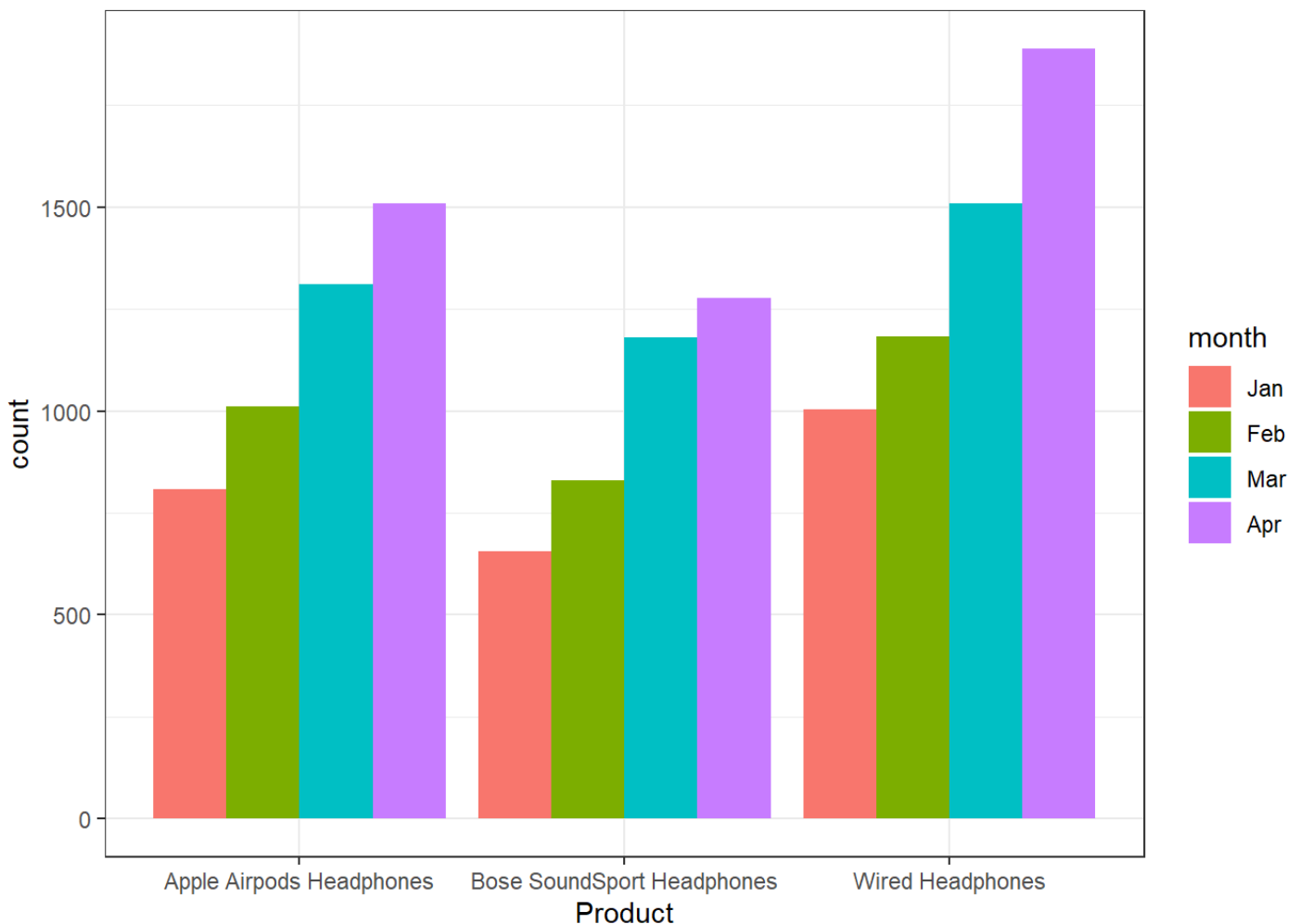
```
## # A tibble: 9 x 2
##   Address          consume
##   <chr>          <dbl>
## 1 " Atlanta"      190.
## 2 " New York City" 188.
## 3 " Portland"     188.
## 4 " Seattle"      188.
## 5 " San Francisco" 187.
## 6 " Dallas"       185.
## 7 " Boston"       182.
## 8 " Los Angeles"   181.
## 9 " Austin"       177.
```

2. Visualization Problem(Using ggplot2)

2.1 헤드폰의 월별 판매량을 그리시오 [10점]

- 상품 중 헤드폰에 해당되는 것들을 filtering 하시오 [5점]
 - 단, 여기선 편의를 위해 Quantity.Ordered는 무시할 것(모두 1로 가정)
- 기본 그림[5점]

```
data %>% filter(Product %in% c("Apple AirPods Headphones", "Bose SoundSport Headphones", "Wired Headphones")) %>% ggplot(aes(Product, fill=month)) + geom_bar(position="dodge") + theme_bw()
```



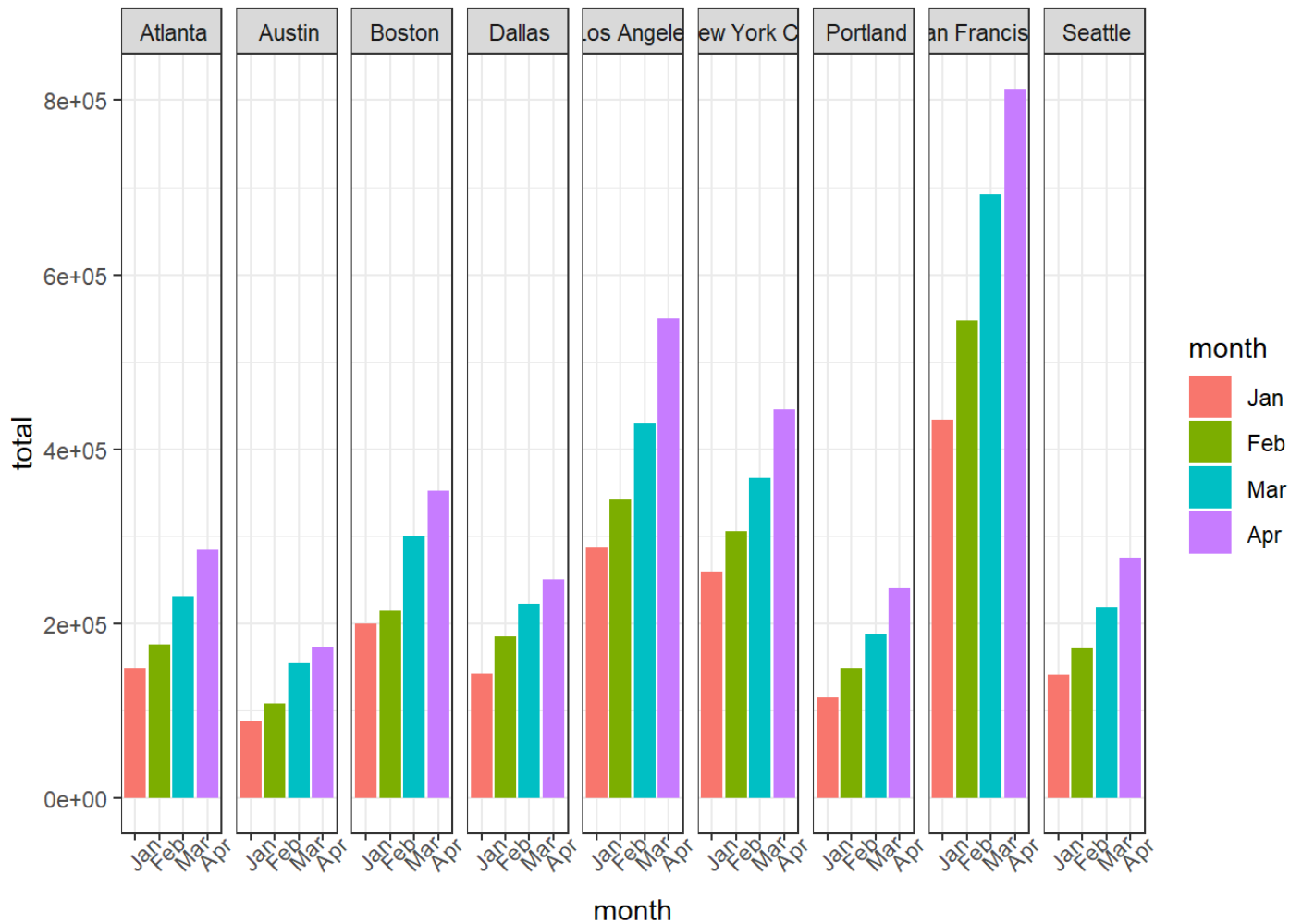
2.2 지역별 총 소비액을 월별로 그리시오 [25점]

- 지역별, 월별 총 소비액을 구하시오. [10점]
 - 힌트: 총 소비액은 Quantity.Ordered와 Price.Each를 곱한 변수를 생성 후 sum()을 활용
- bar plot을 활용[5점]
- 그림과 같이 지역별 그림은 따로 그리시오[10점]
- x축 label을 45도 기울이시오[5점]

```
result=data %>% mutate(consume = Quantity.Ordered*Price.Each) %>% group
_by(Address, month) %>%
  summarise(total = sum(consume))
result
```

```
## # A tibble: 36 x 3
## # Groups:   Address [9]
##   Address      month    total
##   <chr>         <fct>    <dbl>
## 1 " Atlanta" Jan     149160.
## 2 " Atlanta" Feb     176466.
## 3 " Atlanta" Mar     231805.
## 4 " Atlanta" Apr     284443.
## 5 " Austin" Jan       87587.
## 6 " Austin" Feb     108787.
## 7 " Austin" Mar     154546.
## 8 " Austin" Apr     172684.
## 9 " Boston" Jan     199762.
## 10 " Boston" Feb     214793.
## # ... with 26 more rows
```

```
result %>% ggplot(aes(month, total, fill=month)) + geom_bar(stat="ident
ity", position="dodge") + facet_grid(~Address) + theme_bw() + theme(axi
s.text.x = element_text(angle = 45))
```



© 2021 Advice, All Rights Reserved

No part of this contents may be reproduced, copied, modified or adapted, without the prior written consent of the author, unless otherwise indicated for stand-alone materials.