

TEST_03

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

- 필요 패키지 : `dplyr`, `tidyr`, `ggplot2`
 - 필요 데이터 : `Tour_age.csv`, `Tour_gender.csv`, `Tour_purpose.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 데이터를 불러와서 다음과 같이 출력하시오.[10점]

- 데이터를 불러와서 각각 age, gender, purpose로 저장하시오[5점]
- 하나의 데이터 프레임으로 만든 후 다음과 같이 필요한 열만 가져오시오[5점]
 - 반드시 결과물과 같이 출력되도록 할 것(필요한 열만 잘 선택할 것)
- head()로 출력할 것

```
##      date      nation visitor age0.20 age21.30 age31.40 age41.50 age5
1.60 age61
## 1 2019-1      China  392814   36520   108591   103657   48574    4
0893 40998
## 2 2019-1      Japan  206526   18015    57921    34165   39811    3
3857 20330
## 3 2019-1      Taiwan   87954   18888    17927    18595   18862
8169  4566
## 4 2019-1 Hong Kong   35896    3890    11384    7400    5461
4629  2137
## 5 2019-1      Macao    2570     223     1013     762     264
181     92
## 6 2019-1 Phillipines  30473   1436     5051    5486    3140
1831    811
##      male female crewman tourism business official.affairs studying o
thers
## 1 147511 231722   13581  320113     2993             138     8793
60777
## 2  75070 129029    2427  198805     2233             127     785
4576
## 3  30805  56202     947   86393      74             22     180
1285
## 4  12172  22729     995   34653      59             2     90
1092
## 5     748   1787      35    2506      2             0     17
45
## 6  10460   7295   12718  14279     211             161    184
15638
```

1.2 다음과 같이 변수 명을 변경 후 str()을 통해 출력하시오[15점]

- 다음과 같이 열 이름을 변경하시오[5점]
 - 힌트: age로 시작하는 열들의 이름을 살펴볼 것
- date에서 19년도 5월부터 20년도 4월까지만 filtering해서 가져오시오.[5점]
- date는 factor로 나타낼 것.[5점]
 - 힌트: factor의 levels를 변경할 것, labels는 변경할 필요 없음
- 최종 데이터만 str()로 출력할 것

```
## 'data.frame':    720 obs. of  17 variables:
## $ date          : Factor w/ 12 levels "2019-5","2019-6",...: 1 1 1
1 1 1 1 1 1 1 ...
## $ nation        : chr  "China" "Japan" "Taiwan" "Hong Kong" ...
## $ visitor       : int   500413 286273 101779 57026 2909 50569 1893
6 52660 44416 13858 ...
## $ age0~20       : int   17011 20144 10262 5934 182 5084 1662 4348
4331 539 ...
## $ age21~30      : int   138797 71069 21375 13256 924 6991 3677 155
54 10596 1892 ...
## $ age31~40      : int   141104 47535 25044 10654 853 7129 3053 126
07 10812 2822 ...
## $ age41~50      : int   66171 51611 18819 9285 368 5841 2189 7457
6071 1716 ...
## $ age51~60      : int   63504 47821 13791 10340 365 3647 1472 4588
4813 823 ...
## $ age61         : int   58996 45462 11481 6612 216 2491 1001 2623
2431 431 ...
## $ male          : int   187922 104270 32707 17219 758 14645 6964 1
4714 19493 6498 ...
## $ female        : int   297661 179372 68065 38862 2150 16538 6090
32463 19561 1725 ...
## $ crewman       : int   14830 2631 1007 945 1 19386 5882 5483 5362
5635 ...
## $ tourism       : int   413949 279174 100004 55756 2885 29015 1146
4 45671 32528 4483 ...
## $ business      : int   4034 2019 84 56 1 201 154 61 1073 2896 ...
## $ official.affairs: int   534 91 11 2 0 176 152 445 368 99 ...
## $ studying      : int   14003 715 234 147 13 56 75 71 2338 92 ...
## $ others        : int   67893 4274 1446 1065 10 21121 7091 6412 81
09 6288 ...
```

1.3 월별 총 외국인 방문객 수를 구하시오. [10점]

```
## # A tibble: 12 x 2
##   date      total
##   <fct>    <int>
## 1 2019-5    1485684
## 2 2019-6    1476218
## 3 2019-7    1448067
## 4 2019-8    1586299
## 5 2019-9    1459664
## 6 2019-10   1656195
## 7 2019-11   1456429
## 8 2019-12   1456888
## 9 2020-1    1272708
## 10 2020-2     685212
## 11 2020-3      83497
## 12 2020-4      29415
```

1.4 월별, 나이 구간별 외국인 방문객 수를 구하시오. [20점]

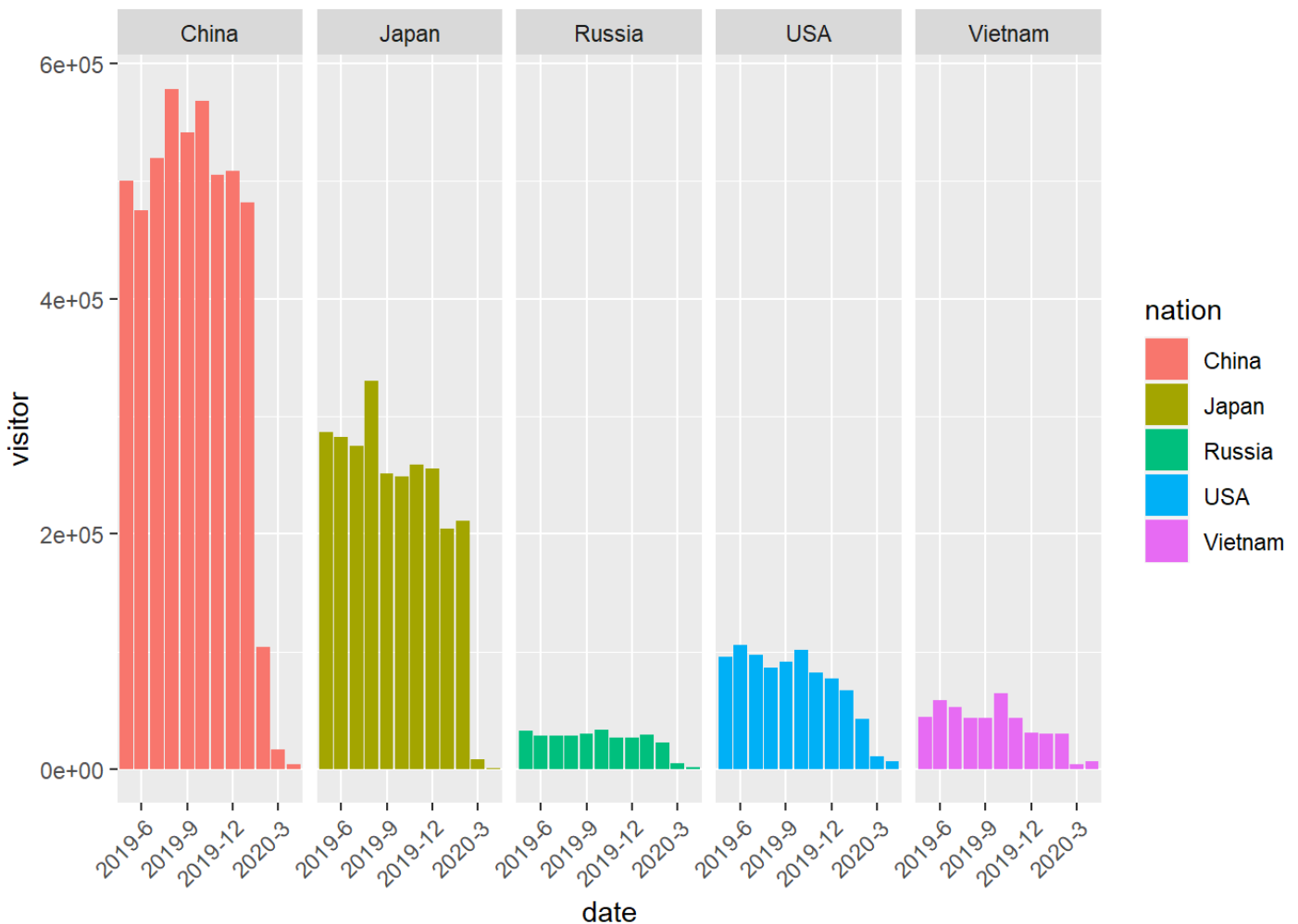
- age로 시작하는 열을 gather를 활용해서 long format으로 바꾸시오. [10점]
- 새롭게 생성된 변수를 활용하여 월별, 나이 구간별 방문객 수를 구하시오. [5점]
- 그룹별 상위 2개만 구하시오. [5점]

```
## # A tibble: 24 x 3
## # Groups:   date [12]
##   date      age      total
##   <fct>    <chr>    <int>
## 1 2019-5 age21~30 355340
## 2 2019-5 age31~40 330804
## 3 2019-6 age21~30 366645
## 4 2019-6 age31~40 321241
## 5 2019-7 age21~30 338386
## 6 2019-7 age31~40 294658
## 7 2019-8 age21~30 388519
## 8 2019-8 age31~40 306222
## 9 2019-9 age21~30 395973
## 10 2019-9 age31~40 334089
## # ... with 14 more rows
```

2. Visualization Problem(Using ggplot2)

2.1 국가별 방문객 수 추이를 그리시오 [25점]

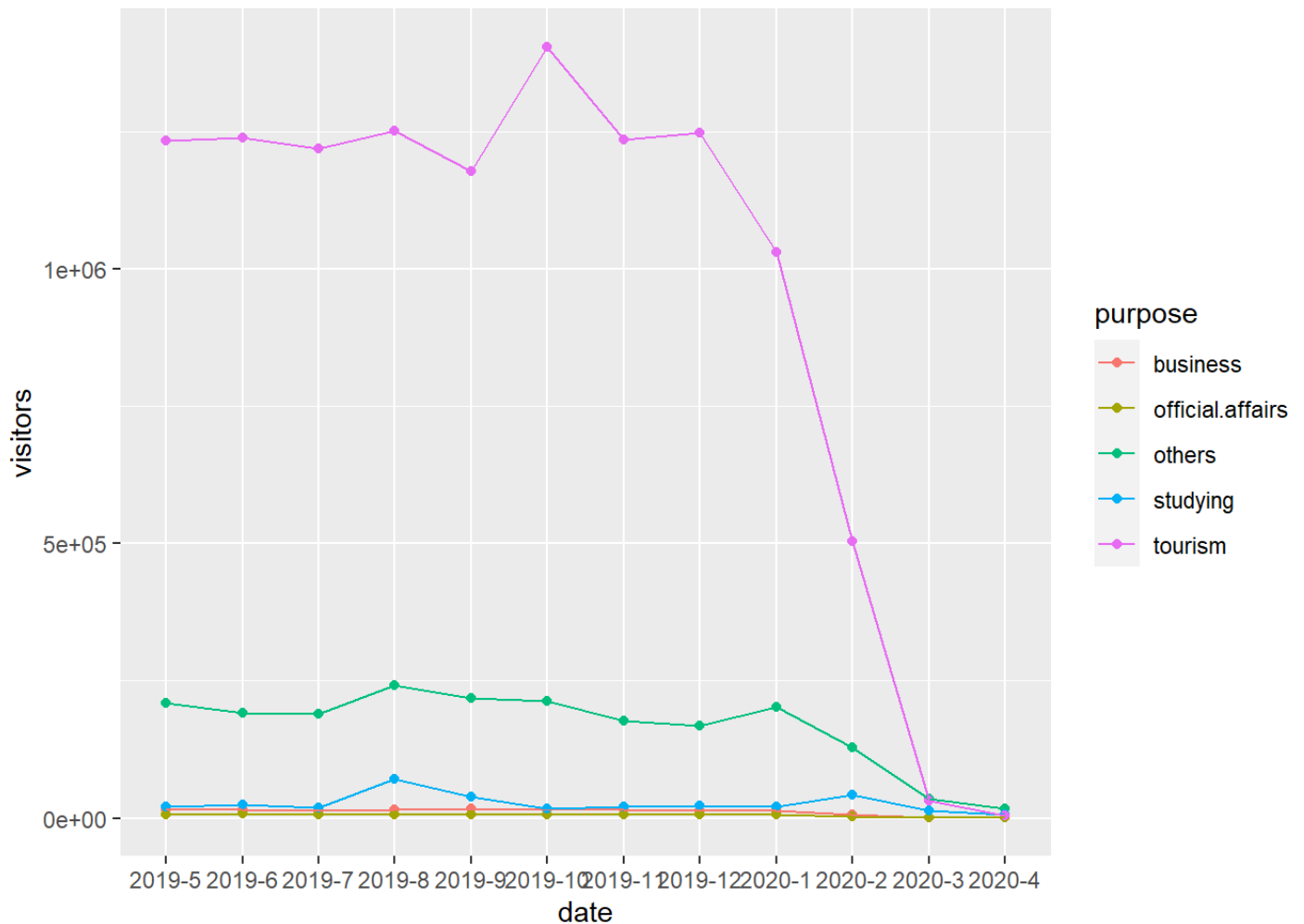
- “China”, “Japan”, “USA”, “Vietnam”, “Russia”에 해당되는 국가만 filtering 하시오.[5점]
- 기본 그림[5점]
- 국가별로 그림을 따로 나타내시오[5점]
- x축 label을 그림과 같이 일부만 표시되게 하시오.[5점]
 - 2019-6, 2019-9, 2019-12, 2020-3만 표시되게 하시오
- x축 label을 45도 기울이시오.[5점]



2.2 방문 목적의 변화를 월별로 그리시오 [20점]

- 방문 목적 5가지를 gather를 활용해서 long format으로 바꾸시오.[5점]
 - 방문 목적: tourism, business, official.affairs, studying, others
- 월별, 방문 목적별 방문객 수를 구하고 head()를 통해서 10개만 나타낼 것.[5점]
- 기본 그림[10점]
 - 선이 제대로 안그려지면 부분점수 없음

```
## # A tibble: 10 x 3
## # Groups:   date [2]
##   date   purpose      visitors
##   <fct> <chr>         <int>
## 1 2019-5 business      15020
## 2 2019-5 official.affairs    6457
## 3 2019-5 others        209943
## 4 2019-5 studying       20702
## 5 2019-5 tourism      1233562
## 6 2019-6 business      15006
## 7 2019-6 official.affairs    7327
## 8 2019-6 others        192227
## 9 2019-6 studying       23818
## 10 2019-6 tourism      1237840
```



© 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.