DAY 2 ASSIGNMENT

Understanding Grouping and Transforming Data

Instruction on the Day 2 assignment

DAY 2 과제는 크게 세 개 파트로 나누어져 있습니다. 첫 번째 파트는 {tidyverse} 패키지의 함수 중 하나인 group_by()에 대한 이해를 돕기 위한 것이며, 두 번째 파트는 {tidyr} 패키지를 활용하여 데이터를 목적에 맞게 전처리하는 방식 중 하나인 형태 변환(transformation)을 수행할 수 있는지 확인하기 위한 것입니다. 마지막으로는 각각의 파트에 제시된 문제를 바탕으로 그 결과를 시각화하여 나타내는 것입니다. 각각의 문제는 RMarkdown 또는 Quarto를 이용하여 PDF 형태로 제출하여 주시기 바랍니다. 혹은 PDF 컴파일에 성공하지 못하였다면, .Rmd나 .qmd 파일의 형태로 제출하셔도 좋습니다.

```
library(tidyverse)
```

Part I. group_by()

{fivethirtyeight} 패키지를 설치하고 bechdel 데이터를 불러오세요.

```
if (!require(fivethirtyeight)) install.packages("fivethirtyeight")
```

```
Loading required package: fivethirtyeight
```

```
Some larger datasets need to be installed separately, like senators and
house_district_forecast. To install these, we recommend you install the
fivethirtyeightdata package by running:
install.packages('fivethirtyeightdata', repos =
'https://fivethirtyeightdata.github.io/drat/', type = 'source')
```

```
library(fivethirtyeight)
data(bechdel)
```

{psych} 패키지의 describe() 함수를 이용하거나 {skmir} 패키지의 skim() 함수를 이용해서 bechel 데이터셋에 대해 간단하게 요약 및 정리하여 서술하세요.

데이터셋 안에 총 몇 개의 관측치가 있는지 계산하세요.

domgross와 intgross 간의 비율을 보여주는 변수, gross_prop를 새로 만들어서 데이터에 저장하세요.

모든 영화의 평균 domgross 값을 계산하세요.

binary 변수는 영화가 심의를 통과했는지 여부를 보여주는 변수입니다. 심의를 통과한 영화와 그렇지 않은 영화 각각의 domgross 평균을 계산하세요. (반드시 group_by()를 사용할 것)

검증을 통과한 영화들만을 포함한 데이터셋, passed를 새롭게 만드세요.

domgross와 intgross의 관계를 산포도로 나타내세요.

clean_test의 카테고리별로 domgross와 intgross 간의 관계를 산포도로 나타내세요.

Part II. {tidyr}: pivot_longer()와 pivot_wider()

```
people <- tribble(
  ~name,             ~names,  ~values,
  #-----------------|--------|------
  "Phillip Woods",   "age",       45,
  "Phillip Woods",   "height",   186,
  "Jessica Cordero", "age",       37,
  "Jessica Cordero", "height",   156
)
```

주어진 데이터셋 people을 넓은 형태로 변환하여 보세요. names에 속한 각 값들이 고유한 변수로, values에 해당하는 값을 갖도록 하세요.

다음의 preg 데이터를 깔끔하게 정돈해보세요. 길게 만들어야 할까요, 아니면 넓게 만들어야 할까요? 변환한다면 어떤 변수들이 추가되거나 사라져야 할까요?

```
preg <- tribble(
  ~pregnant, ~male, ~female,
  "yes",     NA,    10,
  "no",      20,    12
)
```

이번에는 정치학 데이터를 사용해서 분석해보도록 하겠습니다.

```
if (!require(vdemdata)) devtools::install_github("vdeminstitute/vdemdata")
```

```
Loading required package: vdemdata
```

```
vdem_sub <- vdemdata::vdem |>
  dplyr::select(country_name, country_id, year,
                v2x_polyarchy, v2elembaut, v2elembaut_ord,
                e_gdppc)
```

민주주의의 다양성(Varieties of Democracy) 데이터셋의 서브셋인 이 데이터에 포함된 변수들은 다음과 같습니다:

- country_name: 국가명
- country_id: 국가 식별 ID
- year: 연도
- v2x_polyarchy: 선거민주주의 지수
- v2elembaut: 선거관리기관(Election management body, EMB)가 전국선거를 잘 운영하고 관리하기 위한 충분한 인력과 자원을 갖추고 있는지에 대한 전문가 평가를 등간척도로 나타낸 것
- v2elembaut_ord: v2elembaut_ord을 순서형 변수로 나타낸 것

- e_gdppc: 1인당 GDP

더 자세한 변수들의 설명은 V-Dem codebook을 통해 확인하세요.

country_name, country_id, year를 제외한 변수들의 유형이 무엇인지 서술하세요.

v2x_polyarchy의 분포를 그래프로 나타내세요.

직전 연도의 v2x_polyarchy와 당해 년도의 v2x_polyarchy 값의 관계를 그래프로 나타내세요. (group_by와 lag 함수 이용)

v2elembaut와 v2elembaut_ord이 어떻게 다른지 그래프를 통해 보여주세요. 변수의 유형이라고 하는 것이 어떠한 차이를 가져올 수 있는지 자신의 생각을 알려주세요.

e_gdppc을 $1,025 이하일 경우에 "low income", $1,025 초과 $12,375 이하일 경우에는 "middle income", $12,375 초과일 경우에 "high income"을 갖는 분류형 변수 gdppc_bin을 만들고 각각 다음에 해당하는 질문에 답하세요. 단, 데이터에서 e_gdppc 데이터의 단위를 확인하세요.:

2015년에 각 소득 구간에 속한 국가들의 수는 각각 몇 개인가요?

각 소득 구간별 평균 민주주의 지수는 얼마인가요?