data_prep

December 15, 2023

[1]:
```
# Make Jupyter reload library before every execution

%load_ext autoreload
%autoreload 2
```

1 Translation

Each data file is in a different language. We need to translate them into English before concatenation. First, we'll examine the headers

[2]:
```
# We code some function for preprocessing here
import utils
```

[3]:
```
import pandas as pd

# Load the survey data files
file_paths = {
    'English': 'data/en.csv',
    'Bengali': 'data/bengali.csv',
    'Korean': 'data/kr.csv',
    'Vietnamese': 'data/vi.csv'
}

# Read the headers of each file
dfs = {}
headers = {}
for language, file_path in file_paths.items():
    df = pd.read_csv(file_path, parse_dates=True)  # Read only headers
    dfs[language] = df
    headers[language] = df.columns.tolist()

headers
```

[3]:
```
{'English': ['Timestamp',
  'What is your age group?',
  'What is your gender?',
  'What is the highest level of education you have completed?',
  'Which category best describes your occupation?',
  'On average, how many days per week do you exercise for at least 30 (or under
30 mins but high intensity) minutes? ',
  'On average, how many hours per day do you use electronic devices such as
smartphones, laptops, televisions, etc.? ',
  'On average, how much time do you usually spend on phone / computers before
sleep?',
  '(Optional) What is your height in centimeters?',
  '(Optional) What is your weight in kilograms?',
  'On average, what time do you typically go to bed at night?',
  'On average, what time do you typically wake up in the morning?',
  'On average, how long does it take you to fall asleep at night?',
  'On average, how long is your typical daytime nap?',
  'On average, how many hours do you sleep per 24-hour period?',
  'How would you rate your overall sleep quality?',
  'How often do you experience sleep disturbances such as waking up during the
night or having restless sleep?',
  'Do you take any medication to help you sleep?',
  'Your email address'],
 'Bengali': ['Timestamp',
  'Username',
  '       ?*',
  '        ? *',
  '              ? *\n',
  '        ? *',
  '
      ? *',
  '                           ,         ,    ,
    ,    ? ',
  '               /                  ?',
  '(   )         ? ( .  .   )',
  '(   )        ? (    )',
  '                ?',
  '                      ?',
  '                           ?',
  '                      ?           ,         ?',
  '                             ?',
  '                    ? ',
  '                           ,
         ?',
  '                      ?',
  '          '],
 'Korean': ['Timestamp',
  '          ?',
  '        ?',
  '          ?',
  '        ?',
  '            30        ?',
  '      ,   , TV                ?',
  '     /               ?',
  '(   )         ?',
  '(   )          ?',
  '            ?',
  '           ? ',
  '               ?',
  '            ?',
  '24                 ?',
  '             ?',
  '                      ?',
  '       ?',
  '      '],
 'Vietnamese': ['Timestamp',
  'Bạn bao nhiêu tuổi?',
  'Giới tính của bạn? ',
  'Trình độ học vấn cao nhất của bạn? ',
  'Nghề nghiệp của bạn?',
  'Trung bình mỗi tuần bạn tập thể dục ít nhất 30 phút (hoặc dưới 30p nhưng
cường độ cao) bao nhiêu ngày?',
  'Trung bình mỗi ngày bạn dùng các thiết bị điện tử như điện thoại thông minh,
máy tính xách tay, ti vi, v.v. bao nhiêu giờ?',
  'Trung bình bạn dành bao nhiêu thời gian sử dụng điện thoại/máy tính trước khi
đi ngủ? ',
  '(Không bắt buộc) Chiều cao của bạn là bao nhiêu cm?',
  '(Không bắt buộc) Cân nặng của bạn là bao nhiêu kg?',
  'Thông thường bạn đi ngủ vào lúc mấy giờ tối?',
  'Thông thường bạn thức dậy vào lúc mấy giờ sáng?',
  'Thông thường bạn mất khoảng bao lâu để đi vào giấc ngủ đêm?',
  'Thông thường giấc ngủ trưa của bạn kéo dài bao lâu?',
  'Trung bình mỗi ngày bạn ngủ bao nhiêu giờ trong một ngày (24 giờ)?',
  'Bạn đánh giá chất lượng giấc ngủ chung của mình như thế nào?',
  'Giấc ngủ của bạn có hay bị gián đoạn (vd: thức giấc nửa đêm, ngủ không yên)
khong?',
  'Bạn có sử dụng bất kỳ loại thuốc nào để hỗ trợ giấc ngủ không?',
  'Địa chỉ email của bạn']}
```

Let's drop Username and Email columns for privacy

[4]:
```python
for lang, df in dfs.items():
    try:
        # Email column is the last column
        df.drop(labels=df.columns[-1], axis=1, inplace=True)

        # Drop username column if present
        df.drop(columns='Username', inplace=True)
    except Exception as err:
        pass
        # print(err)
```

As you can see, column headers are lengthy, which could make it harder for us to analyze. In the next section, we will translate and shorten them while trying to maintaining the original meaning as close as possible.

1.1 Bengali

1.1.1 Translate Headers to English

[5]:
```
dfs['Bengali'].head(3)
```

[5]:
```
                     Timestamp        ?*         ? *  \
0  2023/11/07 3:26:01 PM GMT+9            -
1  2023/11/07 3:45:02 PM GMT+9            -
2  2023/11/07 3:51:38 PM GMT+9            -

               ? *\n          ? *   \
0                                  :
1                                  :
2                                   :



  ? *  \
0                                             -
1                                             -
2

                              ,        ,     ,
     ,    ?   \
0                                           -
1                                           -
2

                  /                  ?  \
0                                       -
1
2

   (   )         ? ( .   .   )   (  )         ? (    )  \
0                                    152.4                              47.0
1                                      NaN                               NaN
2                                    153.4                              68.0

                  ?                       ?  \
0                        23:00                                 06:30
```

```
1                        00:00                                  08:30
2                        00:30                                  09:30

                             ?  \
0
1                                       -
2

                        ?          ,          ?  \
0                                 ,
1                                 ,
2                               ,

                               ?  \
0                                           -
1
2                                           -

                       ?   \
0                                      2
1                                      4
2                                      3

                             ,
       ?  \
0
1
2

                        ?
0
1
2
```

Column Headers are too lengthy and hard to read. Let's rename them to English

```
[6]: # Define the new column names
new_column_names = {
    'Timestamp': 'Timestamp',
    '       ?*': 'Age Group',
    '        ? *': 'Gender',
    '              ? *\n': 'Education Level',
    '        ? *': 'Occupation',
    '                                               ? *': 'Exercise Days/Week',
    '                            ,        ,    ,     ,    ? ': 'Device Usage␣
 ↪(hrs/day)',
    '               /                  ?': 'Screen Time Before Sleep',
    '(   )         ? ( . .    )': 'Height (cm)',
    '(   )        ? (    )': 'Weight (kg)',
    '                ?': 'Bedtime',
    '                      ?': 'Wake-up Time',
    '                          ?': 'Sleep Onset Time',
    '                     ?          ,          ?': 'Nap Duration',
    '                            ?': 'Sleep Duration (hrs/24hr)',
    '                    ? ': 'Sleep Quality',
    '                          ,                          ?': 'Sleep␣
  ↪Disturbances',
    '                      ?': 'Sleep Medication',
}

# Rename the columns
dfs['Bengali'].rename(columns=new_column_names, inplace=True)

# Show the updated DataFrame
dfs['Bengali'].head()
```

```
[6]:                      Timestamp Age Group Gender            Education Level  \
0  2023/11/07 3:26:01 PM GMT+9      -
1  2023/11/07 3:45:02 PM GMT+9      -
2  2023/11/07 3:51:38 PM GMT+9      -
3  2023/11/07 3:53:05 PM GMT+9      -
4  2023/11/07 4:05:44 PM GMT+9      -

             Occupation Exercise Days/Week Device Usage (hrs/day)  \
0                :             -                  -
1                :             -                  -
2                :
3       /
4                                                 -

  Screen Time Before Sleep  Height (cm)  Weight (kg) Bedtime Wake-up Time  \
0             -             152.4         47.0    23:00        06:30
1                               NaN          NaN    00:00        08:30
2                         153.4         68.0    00:30        09:30
3             -             154.0         59.2    22:50        06:20
4                             171.0         75.0    22:00        06:00

  Sleep Onset Time           Nap Duration Sleep Duration (hrs/24hr)  \
0                ,                           -
1       -          ,
2               ,                             -
3       -        ,
4       -        ,
   Sleep Quality Sleep Disturbances Sleep Medication
0              2
1              4
2              3
3              3
4              4
```

1.1.2 Translate Cell values to English

[7]:
```
bengali_translation_dict = {
    "Age Group": {
        " - ": "16-24",
        " - ": "25-34",
        " - ": "35-44",
        " - ": "45-54",
        " +": "55+",
        "   ": "Other",
    },
    "Gender": {
        "  ": "Male",
        "   ": "Female",
        "   ": "Other"
    },
    "Education Level": {
        "        ": "Master's",
        "        ": "Bachelor's",
        "              ": "Doctorate",
        "       ": "High School"
    },
    "Occupation": {
        "   :": "Other",
        "    /      ": "Professional/Office Worker",
        "  ": "Student",
        "   ": "Unemployed",
        "    (   ,           )": "Service"
    },
    "Exercise Days/Week": {
        " -  ": "1-2 Days",
        " -  ": "3-4 Days",
        "   ": "0 Days",
        "         ": "5+ Days"
    },
    "Device Usage (hrs/day)": {
        " -   ": "1-3 Hours",
        " -   ": "4-6 Hours",
        "          ": "7+ Hours"
    },
    "Screen Time Before Sleep": {
        "      -     ": "30-60 Minutes",
        "        ": "<30 Minutes",
        "             ": "2+ Hours",
        " -    ": "1-2 Hours"
    },
    "Sleep Onset Time": {
        "         ": ">60 Minutes",
        " -     ": "15-30 Minutes",
        "        ": "<15 Minutes",
        " -     ": "30-60 Minutes"
    },
    "Nap Duration": {
        "  ,         ": "<30 Minutes",
        " ,            ": "No Nap",
        "  ,  -     ": "30-60 Minutes",
        "  ,          ": ">90 Minutes"
    },
    "Sleep Duration (hrs/24hr)": {
        " -    ": "4-6 Hours",
        "            ": "6+ Hours"
    },
    "Sleep Disturbances": {
        "      ": "Often",
        "       ": "Sometimes",
        "      ": "Rarely",
        "     ": "Never"
    },
    "Sleep Medication": {
        " ": "No",
        "  ": "Yes"
    }
}


utils.translate_cells(dfs["Bengali"], bengali_translation_dict)
dfs["Bengali"]["Language"] = "Bengali"
dfs["Bengali"].sample(5)
```

[7]:

```
                        Timestamp Age Group  Gender Education Level  Occupation  \
14  2023/11/07 6:23:15 PM GMT+9     16-24    Male        Master's     Student
5   2023/11/07 4:38:46 PM GMT+9     25-34  Female        Master's  Unemployed
12  2023/11/07 6:19:23 PM GMT+9     45-54  Female     High School     Student
9   2023/11/07 5:49:23 PM GMT+9     25-34  Female        Master's     Student
0   2023/11/07 3:26:01 PM GMT+9     25-34  Female        Master's       Other

   Exercise Days/Week Device Usage (hrs/day) Screen Time Before Sleep  \
14           1-2 Days              4-6 Hours            30-60 Minutes
5            1-2 Days              4-6 Hours                 2+ Hours
12           1-2 Days              4-6 Hours            30-60 Minutes
9              0 Days              1-3 Hours              <30 Minutes
0            1-2 Days              1-3 Hours            30-60 Minutes

    Height (cm)  Weight (kg) Bedtime Wake-up Time Sleep Onset Time  \
14         74.0         55.0   00:00        08:00    30-60 Minutes
5           NaN          NaN   02:00        09:00      >60 Minutes
12          NaN         43.0   00:00        06:00    15-30 Minutes
9         160.0         53.0   12:00        07:00    15-30 Minutes
0         152.4         47.0   23:00        06:30      >60 Minutes

     Nap Duration Sleep Duration (hrs/24hr)  Sleep Quality Sleep Disturbances  \
14  30-60 Minutes                  6+ Hours              3             Rarely
5          No Nap                 4-6 Hours              4              Never
12  30-60 Minutes                 4-6 Hours              3              Never
9   30-60 Minutes                  6+ Hours              3             Rarely
0     <30 Minutes                 4-6 Hours              2              Often

   Sleep Medication Language
14               No  Bengali
5                No  Bengali
12               No  Bengali
9                No  Bengali
0                No  Bengali
```

Now, the Bengali survey data is completely translated. Now, we'll do the same thing other the remaining languages.

1.2 Vietnamese

1.2.1 Before

[8]:
```
dfs['Vietnamese'].sample(5)
```

[8]:
```
                       Timestamp Bạn bao nhiêu tuổi? Giới tính của bạn?   \
5   2023/11/07 12:07:55 PM GMT+9               16-24                 Nam
15   2023/11/08 9:06:41 AM GMT+9               25-34                  Nữ
11   2023/11/07 5:18:05 PM GMT+9               25-34                 Nam
2   2023/11/07 11:48:19 AM GMT+9               25-34                  Nữ
0   2023/11/07 11:38:56 AM GMT+9               25-34                  Nữ

   Trình độ học vấn cao nhất của bạn?      Nghề nghiệp của bạn?  \
5                                  THPT     Học sinh / Sinh viên
15                              Tiến sĩ                Reseacher
11                              Thạc sĩ  Chuyên nghiệp/văn phòng
2                               Thạc sĩ                Giáo viên
```

```
0                              Thạc sĩ  Chuyên nghiệp/văn phòng

   Trung bình mỗi tuần bạn tập thể dục ít nhất 30 phút (hoặc dưới 30p nhưng
cường độ cao) bao nhiêu ngày?  \
5                                      5 ngày trở lên
15                                             0 ngày
11                                           1-2 ngày
2                                              0 ngày
0                                            3-4 ngày

   Trung bình mỗi ngày bạn dùng các thiết bị điện tử như điện thoại thông minh,
máy tính xách tay, ti vi, v.v. bao nhiêu giờ?  \
5                                       7 giờ trở lên
15                                            4-6 giờ
11                                            1-3 giờ
2                                             4-6 giờ
0                                             4-6 giờ

   Trung bình bạn dành bao nhiêu thời gian sử dụng điện thoại/máy tính trước khi
đi ngủ?   \
5                                      Ít hơn 30 phút
15                                    30 phút - 1 giờ
11                                     Ít hơn 30 phút
2                                             1-2 giờ
0                                             1-2 giờ

    (Không bắt buộc) Chiều cao của bạn là bao nhiêu cm?  \
5                                                 NaN
15                                              154.0
11                                              165.0
2                                               160.0
0                                                 NaN

    (Không bắt buộc) Cân nặng của bạn là bao nhiêu kg?  \
5                                                 NaN
15                                               44.0
11                                               60.0
2                                                53.0
0                                                 NaN

   Thông thường bạn đi ngủ vào lúc mấy giờ tối?  \
5                                           23:00
15                                          01:00
11                                          12:00
2                                           01:00
0                                           11:00
```

```
   Thông thường bạn thức dậy vào lúc mấy giờ sáng?  \
5                                            07:00
15                                           08:00
11                                           08:00
2                                            06:00
0                                            06:00

   Thông thường bạn mất khoảng bao lâu để đi vào giấc ngủ đêm?  \
5                                          15-30 phút
15                                         15-30 phút
11                                         15-30 phút
2                                          15-30 phút
0                                      Ít hơn 15 phút

   Thông thường giấc ngủ trưa của bạn kéo dài bao lâu?  \
5                                          30-60 phút
15                                 Tôi không ngủ trưa
11                                 Tôi không ngủ trưa
2                                  Tôi không ngủ trưa
0                                  Tôi không ngủ trưa

   Trung bình mỗi ngày bạn ngủ bao nhiêu giờ trong một ngày (24 giờ)?  \
5                                           Hơn 6 giờ
15                                            4-6 giờ
11                                          Hơn 6 giờ
2                                             4-6 giờ
0                                           Hơn 6 giờ

    Bạn đánh giá chất lượng giấc ngủ chung của mình như thế nào?  \
5                                                   4
15                                                  2
11                                                  3
2                                                   5
0                                                   4

   Giấc ngủ của bạn có hay bị gián đoạn (vd: thức giấc nửa đêm, ngủ không yên)
khong?  \
5                                            Hiếm khi
15                                       Thỉnh thoảng
11                                       Thỉnh thoảng
2                                       Không bao giờ
0                                            Hiếm khi

   Bạn có sử dụng bất kỳ loại thuốc nào để hỗ trợ giấc ngủ không?
5                                               Không
15                                              Không
11                                              Không
```

```
2                                                Không
0                                                Không
```

1.2.2 After

[9]:

```
# Translate columns to English
vi_headers_dict = utils.read_json('translation/vi_header.json')

dfs['Vietnamese'].rename(columns=vi_headers_dict, inplace=True)

# Translate cell values to English
vi_cells_dict = utils.read_json('translation/vi_val.json')
utils.translate_cells(dfs["Vietnamese"], vi_cells_dict)

dfs["Vietnamese"]["Language"] = "Vietnamese"
dfs['Vietnamese'].sample(5)
```

[9]:

```
                       Timestamp Age Group  Gender Education Level  \
4   2023/11/07 12:03:04 PM GMT+9     25-34    Male      Bachelor's
5   2023/11/07 12:07:55 PM GMT+9     16-24    Male     High School
3   2023/11/07 11:58:10 AM GMT+9     16-24    Male     High School
12   2023/11/07 5:18:07 PM GMT+9     25-34  Female       Doctorate
0   2023/11/07 11:38:56 AM GMT+9     25-34  Female        Master's

                    Occupation Exercise Days/Week Device Usage (hrs/day)  \
4   Professional/Office Worker             0 Days               7+ Hours
5                      Student            5+ Days               7+ Hours
3                      Student             0 Days               7+ Hours
12                     Student           1-2 Days              4-6 Hours
0   Professional/Office Worker           3-4 Days              4-6 Hours

   Screen Time Before Sleep  Height (cm)  Weight (kg) Bedtime Wake-up Time  \
4                  2+ Hours        168.0         60.0   23:00        07:00
5               <30 Minutes          NaN          NaN   23:00        07:00
3               <30 Minutes        175.0         85.0   23:30        05:30
12              <30 Minutes          NaN          NaN   21:30        06:30
0                 1-2 Hours          NaN          NaN   11:00        06:00

   Sleep Onset Time   Nap Duration Sleep Duration (hrs/24hr)  Sleep Quality  \
4       <15 Minutes  30-60 Minutes                  6+ Hours              5
5     15-30 Minutes  30-60 Minutes                  6+ Hours              4
3       <15 Minutes         No Nap                  6+ Hours              5
12    15-30 Minutes         No Nap                  6+ Hours              3
0       <15 Minutes         No Nap                  6+ Hours              4

   Sleep Disturbances Sleep Medication    Language
4              Rarely               No  Vietnamese
```

```
5              Rarely                   No  Vietnamese
3              Rarely                   No  Vietnamese
12             Rarely                   No  Vietnamese
0              Rarely                   No  Vietnamese
```

1.3 English

1.3.1 Before

[10]:

```
dfs["English"].sample(5)
```

[10]:

```
                       Timestamp What is your age group? What is your gender?  \
43   2023/11/08 9:07:21 PM GMT+9                   35-44                Female
18   2023/11/07 8:13:13 PM GMT+9                   25-34                Female
30  2023/11/08 10:13:25 AM GMT+9                   35-44                Female
53   2023/11/09 1:15:14 PM GMT+9                   25-34                  Male
24  2023/11/07 10:40:05 PM GMT+9                   25-34                  Male

   What is the highest level of education you have completed?  \
43                                  Bachelor's degree
18                                    Master's degree
30                                    Master's degree
53                                    Master's degree
24                                    Master's degree

   Which category best describes your occupation?  \
43          Service (retail, food service, etc.)
18                       Professional/office job
30                       Professional/office job
53                                       Student
24                                       Student

   On average, how many days per week do you exercise for at least 30 (or under
30 mins but high intensity) minutes?   \
43                                           3-4 days
18                                           1-2 days
30                                             0 days
53                                           3-4 days
24                                           1-2 days

   On average, how many hours per day do you use electronic devices such as
smartphones, laptops, televisions, etc.?   \
43                                          4-6 hours
18                                          1-3 hours
30                                    7 or more hours
53                                          4-6 hours
24                                    7 or more hours
```

```
   On average, how much time do you usually spend on phone / computers before
sleep?  \
43                                          1-2 hours
18                                30 minutes - 1 hour
30                                  More than 2 hours
53                                30 minutes - 1 hour
24                                  More than 2 hours

    (Optional) What is your height in centimeters?  \
43                                           156.0
18                                           154.0
30                                           150.0
53                                           160.0
24                                           160.0

    (Optional) What is your weight in kilograms?  \
43                                          96.0
18                                          63.0
30                                          51.0
53                                          55.0
24                                          64.0

   On average, what time do you typically go to bed at night?  \
43                                              23:00
18                                              00:30
30                                              02:00
53                                              02:00
24                                              01:00

   On average, what time do you typically wake up in the morning?  \
43                                              05:00
18                                              07:00
30                                              04:30
53                                              09:00
24                                              09:00

   On average, how long does it take you to fall asleep at night?  \
43                                      15-30 minutes
18                                      30-60 minutes
30                               More than 60 minutes
53                               Less than 15 minutes
24                               Less than 15 minutes

   On average, how long is your typical daytime nap?  \
43                   No, I do not nap during the day
18                         Yes, More than 90 minutes
```

```
30                                Yes, 30-60 minutes
53                         Yes, less than 30 minutes
24                   No, I do not nap during the day

   On average, how many hours do you sleep per 24-hour period?  \
43                                  More than 6 hours
18                                  More than 6 hours
30                                          4-6 hours
53                                  More than 6 hours
24                                          4-6 hours

    How would you rate your overall sleep quality?  \
43                                               4
18                                               4
30                                               3
53                                               4
24                                               3

   How often do you experience sleep disturbances such as waking up during the
night or having restless sleep?  \
43                                          Sometimes
18                                          Sometimes
30                                         Frequently
53                                             Rarely
24                                              Never

   Do you take any medication to help you sleep?
43                                            No
18                                            No
30                                            No
53                                            No
24                                            No
```

1.3.2 After

Even though the original survey data is in English, there are two problems: - Column headers contain long text, which decreases readablity. - Cell values also include long text, and unstandardized.

[11]:
```python
# Shorten column header texts
en_headers_dict = utils.read_json('translation/en_header.json')
dfs['English'].rename(columns=en_headers_dict, inplace=True)

# Translate cell values to English
en_cells_dict = utils.read_json('translation/en_val.json')
utils.translate_cells(dfs["English"], en_cells_dict)
dfs["English"]["Language"] = "English"
dfs["English"].sample(5)
```

[11]:

```
                       Timestamp Age Group  Gender Education Level  \
66  2023/11/10 10:15:46 PM GMT+9     25-34  Female      Bachelor's
58   2023/11/09 1:45:22 PM GMT+9     35-44    Male      Bachelor's
56   2023/11/09 1:42:38 PM GMT+9     45-54    Male      Bachelor's
30  2023/11/08 10:13:25 AM GMT+9     35-44  Female        Master's
12   2023/11/07 5:33:48 PM GMT+9     25-34  Female      Bachelor's

                    Occupation Exercise Days/Week Device Usage (hrs/day)  \
66  Professional/Office Worker           1-2 Days              4-6 Hours
58  Professional/Office Worker           1-2 Days               7+ Hours
56  Professional/Office Worker            5+ Days              4-6 Hours
30  Professional/Office Worker             0 Days               7+ Hours
12                     Student           1-2 Days               7+ Hours

   Screen Time Before Sleep  Height (cm)  Weight (kg) Bedtime Wake-up Time  \
66            30-60 Minutes        160.0         63.0   22:30        07:30
58                1-2 Hours        155.0         72.0   23:00        06:00
56                1-2 Hours          NaN         75.0   23:30        05:00
30                 2+ Hours        150.0         51.0   02:00        04:30
12            30-60 Minutes        155.0         55.0   21:30        05:00

   Sleep Onset Time   Nap Duration Sleep Duration (hrs/24hr)  Sleep Quality  \
66    30-60 Minutes    <30 Minutes                  6+ Hours              3
58    15-30 Minutes         No Nap                  6+ Hours              3
56      <15 Minutes  30-60 Minutes                 4-6 Hours              3
30      >60 Minutes  30-60 Minutes                 4-6 Hours              3
12    15-30 Minutes    <30 Minutes                 4-6 Hours              4

   Sleep Disturbances Sleep Medication Language
66          Sometimes               No  English
58             Rarely               No  English
56          Sometimes               No  English
30         Frequently               No  English
12          Sometimes               No  English
```

1.4 Korean

[12]:

```python
# Shorten column header texts
kr_headers_dict = utils.read_json('translation/kr_header.json')
dfs['Korean'].rename(columns=kr_headers_dict, inplace=True)

# Translate cell values to English
kr_cells_dict = utils.read_json('translation/kr_val.json')
utils.translate_cells(dfs["Korean"], kr_cells_dict)
dfs["Korean"]["Language"] = "Korean"
dfs["Korean"]
```

[12]:
```
                      Timestamp Age Group  Gender Education Level Occupation  \
0  2023/11/07 11:29:18 AM GMT+9     16-24  Female     High School    Student

  Exercise Days/Week Device Usage (hrs/day) Screen Time Before Sleep  \
0           1-2 Days               7+ Hours                1-2 Hours

   Height (cm)  Weight (kg) Bedtime Wake-up Time Sleep Onset Time  \
0          167           60   01:00        08:30      <15 Minutes

  Nap Duration Sleep Duration (hrs/24hr)  Sleep Quality Sleep Disturbances  \
0       No Nap                  6+ Hours              4          Sometimes

  Sleep Medication Language
0               No   Korean
```

2 Merge

[13]:
```
df_merge = pd.concat(dfs.values())
df_merge.reset_index(inplace=True, drop=True)
df_merge.sample(5)
```

[13]:
```
                       Timestamp Age Group  Gender Education Level  \
17   2023/11/07 7:32:06 PM GMT+9     16-24  Female      Bachelor's
97  2023/11/07 12:07:55 PM GMT+9     16-24    Male     High School
58   2023/11/09 1:45:22 PM GMT+9     35-44    Male      Bachelor's
8    2023/11/07 5:12:43 PM GMT+9     16-24  Female      Bachelor's
78   2023/11/07 5:52:10 PM GMT+9     25-34  Female        Master's

                    Occupation Exercise Days/Week Device Usage (hrs/day)  \
17                     Student            5+ Days              4-6 Hours
97                     Student            5+ Days               7+ Hours
58  Professional/Office Worker           1-2 Days               7+ Hours
8                      Student           3-4 Days               7+ Hours
78                       Other             0 Days              1-3 Hours

   Screen Time Before Sleep  Height (cm)  Weight (kg) Bedtime Wake-up Time  \
17              <30 Minutes        162.0         46.0   11:00        08:30
97              <30 Minutes          NaN          NaN   23:00        07:00
58                1-2 Hours        155.0         72.0   23:00        06:00
8                 1-2 Hours        155.0         45.0   10:00        05:00
78              <30 Minutes          NaN          NaN   11:00        08:30

   Sleep Onset Time   Nap Duration Sleep Duration (hrs/24hr)  Sleep Quality  \
17    15-30 Minutes  30-60 Minutes                  6+ Hours              3
97    15-30 Minutes  30-60 Minutes                  6+ Hours              4
58    15-30 Minutes         No Nap                  6+ Hours              3
```

```
8     30-60 Minutes  30-60 Minutes                4-6 Hours              3
78    15-30 Minutes    <30 Minutes                4-6 Hours              4

   Sleep Disturbances Sleep Medication    Language
17             Rarely               No     English
97             Rarely               No  Vietnamese
58             Rarely               No     English
8           Sometimes               No     English
78          Sometimes               No     Bengali
```

[14]:

```
df_merge.describe()
```

[14]:

```
       Height (cm)  Weight (kg)  Sleep Quality
count    89.000000    92.000000     108.000000
mean    157.569551    67.415217       3.444444
std      30.981275    12.798085       0.824092
min       5.110000    43.000000       2.000000
25%     155.000000    59.800000       3.000000
50%     165.000000    68.000000       3.000000
75%     171.000000    75.000000       4.000000
max     185.000000   100.000000       5.000000
```

There are twos outlier where height are under 100 (cm). To be safe, I'll replace these values with NaN

[15]:

```
import numpy as np
df_merge['Height (cm)'] = df_merge['Height (cm)'].apply(lambda x: np.nan if x <␣
 ↪100 else x)
df_merge.describe()
```

[15]:

```
       Height (cm)  Weight (kg)  Sleep Quality
count    83.000000    92.000000     108.000000
mean    165.305542    67.415217       3.444444
std       8.321679    12.798085       0.824092
min     150.000000    43.000000       2.000000
25%     160.000000    59.800000       3.000000
50%     167.000000    68.000000       3.000000
75%     171.000000    75.000000       4.000000
max     185.000000   100.000000       5.000000
```

Let's add BMI index, which could be a helpful indicator for health

[16]:

```
def calculate_bmi(weight, height_cm):
    if pd.notnull(weight) and pd.notnull(height_cm):
        height_m = height_cm / 100.0  # Convert height from cm to m
        bmi = weight / (height_m ** 2)
        return round(bmi,1)
    else:
        return None  # Return None if weight or height is missing

df_merge['BMI'] = df_merge.apply(lambda row: calculate_bmi(row['Weight (kg)'],␣
 ↪row['Height (cm)']), axis=1)
df_merge.describe()
```

[16]:

```
       Height (cm)  Weight (kg)  Sleep Quality        BMI
count    83.000000    92.000000     108.000000  80.000000
mean    165.305542    67.415217       3.444444  24.552500
std       8.321679    12.798085       0.824092   4.245503
min     150.000000    43.000000       2.000000  17.500000
25%     160.000000    59.800000       3.000000  21.500000
50%     167.000000    68.000000       3.000000  23.550000
75%     171.000000    75.000000       4.000000  26.600000
max     185.000000   100.000000       5.000000  39.400000
```

Calculate Sleep Duration to compare with self-reported value in later analysis

[17]:

```
from datetime import datetime, timedelta

# Function to parse time considering the day might change over midnight
def parse_time(time_str):
    # Assuming the time format is "HH:MM"
    return datetime.strptime(time_str, "%H:%M").time()

# Function to calculate sleep duration
def calculate_sleep_duration(bedtime_str, wakeup_str):
    bedtime = parse_time(bedtime_str)
    wakeup = parse_time(wakeup_str)

    # Convert to datetime objects
    bedtime_dt = datetime.combine(datetime.today(), bedtime)
    wakeup_dt = datetime.combine(datetime.today(), wakeup)

    # If bedtime is later than wakeup time, assume sleeping past midnight
    if bedtime_dt > wakeup_dt:
        wakeup_dt += timedelta(days=1)

    # Calculate the duration and convert to hours
    duration = wakeup_dt - bedtime_dt
    return round((duration.total_seconds() / 3600 ),2) # convert seconds to␣
 ↪hours

# Apply the function to each row in the DataFrame
df_merge['Calculated Night Sleep Duration'] = df_merge.apply(lambda x:␣
 ↪calculate_sleep_duration(x['Bedtime'], x['Wake-up Time']), axis=1)
df_merge.describe()
```

```
[17]:        Height (cm)  Weight (kg)  Sleep Quality        BMI  \
      count    83.000000    92.000000     108.000000  80.000000
      mean    165.305542    67.415217       3.444444  24.552500
      std       8.321679    12.798085       0.824092   4.245503
      min     150.000000    43.000000       2.000000  17.500000
      25%     160.000000    59.800000       3.000000  21.500000
      50%     167.000000    68.000000       3.000000  23.550000
      75%     171.000000    75.000000       4.000000  26.600000
      max     185.000000   100.000000       5.000000  39.400000

             Calculated Night Sleep Duration
      count                       108.000000
      mean                          9.793981
      std                           5.497084
      min                           1.670000
      25%                           6.500000
      50%                           7.500000
      75%                           9.000000
      max                          23.950000
```

Possible Data Error: Calculate sleep duration have some values over 14 hours

```
[18]: df_merge.sort_values(by="Calculated Night Sleep Duration",␣
        ↪ascending=False)[['Bedtime', 'Wake-up Time', 'Calculated Night Sleep␣
        ↪Duration']]
```

```
[18]:    Bedtime Wake-up Time  Calculated Night Sleep Duration
      23    19:23        19:20                            23.95
      19    00:12        00:07                            23.92
      17    11:00        08:30                            21.50
      78    11:00        08:30                            21.50
      5     11:00        08:00                            21.00
      ..      …            …                               …
      4     01:00        05:30                             4.50
      98    01:00        05:00                             4.00
      20    01:00        05:00                             4.00
      30    02:00        04:30                             2.50
      86    02:20        04:00                             1.67

      [108 rows x 3 columns]
```

Some people may confuse AM / PM and 24-hour format. For example:

```
[19]: df_merge[df_merge['Calculated Night Sleep Duration'] > 12]["Bedtime"].unique()
```

```
[19]: array(['12:30', '11:00', '10:00', '00:12', '04:00', '19:23', '12:00',
             '10:30', '09:30', '13:30', '13:00', '12:15'], dtype=object)
```

The survey is conducted using 24-hour format. So, if a person sleeps at 11PM, he/she should enter 23:00 instead of 11:00, which could be case there. I'll create a function to add 12 hours to their input, except for the "4:00" bed-time above.

```
[20]: def fix_bedtime(bedtime_str):
          bedtime = parse_time(bedtime_str)

          # Convert to datetime objects
          bedtime_dt = datetime.combine(datetime.today(), bedtime)

          # If bedtime is later than wakeup time, assume sleeping past midnight
          if bedtime_dt > datetime.combine(datetime.today(), parse_time("9:00")) and␣
       ↪bedtime_dt < datetime.combine(datetime.today(), parse_time("19:00")):
              bedtime_dt += timedelta(hours=12)

          return bedtime_dt.strftime("%H:%M")

      # Fix the bedtime
      df_merge['Bedtime'] = df_merge.apply(lambda x: fix_bedtime(x['Bedtime']),␣
       ↪axis=1)
      # # Recalculate bedtime
      df_merge['Calculated Night Sleep Duration'] = df_merge.apply(lambda x:␣
       ↪calculate_sleep_duration(x['Bedtime'], x['Wake-up Time']), axis=1)
      # # Check for outliers
      df_merge.sort_values(by="Calculated Night Sleep Duration",␣
       ↪ascending=False)[['Bedtime', 'Wake-up Time', 'Calculated Night Sleep␣
       ↪Duration']]
```

```
[20]:    Bedtime Wake-up Time  Calculated Night Sleep Duration
      23   19:23        19:20                            23.95
      19   00:12        00:07                            23.92
      21   04:00        23:00                            19.00
      74   22:00        07:45                             9.75
      17   23:00        08:30                             9.50
      ..     …            …                                …
      57   00:00        04:00                             4.00
      98   01:00        05:00                             4.00
      20   01:00        05:00                             4.00
      30   02:00        04:30                             2.50
      86   02:20        04:00                             1.67

      [108 rows x 3 columns]
```

There are still 3 strange values at the top, while the majority is now normal.

Let's NaN those values and save the data to csv

[21]:

```
df_merge["Calculated Night Sleep Duration"] = df_merge["Calculated Night Sleep Duration"].apply(lambda x: x if x < 10 else np.nan)
df_merge.to_csv('data/all.csv', index=False)
```