

Chapter 3 - Ex1: Tidying Data

Câu 1:

Cho dữ liệu BMI.xlsx. Bộ dữ liệu này ghi lại BMI của một số quốc gia trong một số năm.

- Đọc dữ liệu
- Xem xét vấn đề về dữ liệu cần khắc phục
- Chuẩn lại dữ liệu để khắc phục vấn đề trên

Đặt yêu cầu ngược lại là cần phải tạo pivot table để xem thống kê theo country, year.

- Hãy chuyển dữ liệu mới làm ở trên về dạng thống kê

Câu 2:

Cho dữ liệu student.xlsx. Bộ dữ liệu này ghi lại điểm các môn của sinh viên

- Đọc dữ liệu
- Xem xét vấn đề về dữ liệu cần khắc phục
- Chuẩn lại dữ liệu để khắc phục vấn đề trên

Đặt yêu cầu ngược lại là cần phải tạo pivot table để xem thống kê theo từng student và các subject.

- Hãy chuyển dữ liệu mới làm ở trên về dạng thống kê

Câu 1: Gợi ý

```
In [1]: import pandas as pd
```

```
In [2]: df = pd.read_excel("BMI.xlsx")
df.head()
```

```
Out[2]:
```

| | Country | Y1980 | Y1981 | Y1982 | Y1983 |
|---|-------------|-------|-------|-------|-------|
| 0 | Afghanistan | 21.48 | 21.46 | 21.45 | 21.44 |
| 1 | Albania | 25.22 | 25.24 | 25.26 | 25.27 |
| 2 | Algeria | 22.26 | 22.35 | 22.44 | 22.52 |

Vấn đề cần khắc phục:

- Tên cột chứa giá trị thay vì chứa biến

```
In [3]: # Melt df into new dataframe: df_melted
df_melted = pd.melt(df, id_vars=['Country'],
                    var_name="year",
                    value_name="bmi")
```

```
In [4]: df_melted
```

```
Out[4]:
```

| | Country | year | bmi |
|----|-------------|-------|-------|
| 0 | Afghanistan | Y1980 | 21.48 |
| 1 | Albania | Y1980 | 25.22 |
| 2 | Algeria | Y1980 | 22.26 |
| 3 | Afghanistan | Y1981 | 21.46 |
| 4 | Albania | Y1981 | 25.24 |
| 5 | Algeria | Y1981 | 22.35 |
| 6 | Afghanistan | Y1982 | 21.45 |
| 7 | Albania | Y1982 | 25.26 |
| 8 | Algeria | Y1982 | 22.44 |
| 9 | Afghanistan | Y1983 | 21.44 |
| 10 | Albania | Y1983 | 25.27 |
| 11 | Algeria | Y1983 | 22.52 |

```
In [5]: df_melted.year = df_melted.year.map(lambda x: int(x[1:]))
```

```
In [6]: df_melted
```

```
Out[6]:
```

| | Country | year | bmi |
|----|-------------|------|-------|
| 0 | Afghanistan | 1980 | 21.48 |
| 1 | Albania | 1980 | 25.22 |
| 2 | Algeria | 1980 | 22.26 |
| 3 | Afghanistan | 1981 | 21.46 |
| 4 | Albania | 1981 | 25.24 |
| 5 | Algeria | 1981 | 22.35 |
| 6 | Afghanistan | 1982 | 21.45 |
| 7 | Albania | 1982 | 25.26 |
| 8 | Algeria | 1982 | 22.44 |
| 9 | Afghanistan | 1983 | 21.44 |
| 10 | Albania | 1983 | 25.27 |
| 11 | Algeria | 1983 | 22.52 |

```
In [7]: df_pivot = df_melted.pivot(index='Country', columns='year', values='bmi')
```

```
In [8]: df_pivot
```

```
Out[8]:
```

| | year | 1980 | 1981 | 1982 | 1983 |
|-------------|-------|-------|-------|-------|------|
| Country | | | | | |
| Afghanistan | 21.48 | 21.46 | 21.45 | 21.44 | |
| Albania | 25.22 | 25.24 | 25.26 | 25.27 | |
| Algeria | 22.26 | 22.35 | 22.44 | 22.52 | |

Câu 2

- Các bạn tự làm nhé.

```
In [9]: import pandas as pd

df = pd.read_excel("student.xlsx")
df.head()
```

```
Out[9]:
```

| | Student | Physics | Chemistry | English | Math |
|---|---------|---------|-----------|---------|------|
| 0 | John | 78 | 79 | 56 | 95 |
| 1 | Alice | 58 | 72 | 91 | 81 |
| 2 | Rachel | 22 | 61 | 88 | 64 |
| 3 | Tom | 78 | 89 | 56 | 83 |

```
In [10]: # Melt df into new dataframe: df_melted
df_melted = pd.melt(df, id_vars=['Student'],
                    var_name="Subject",
                    value_name="Marks")
```

```
In [11]: df_melted.head()
```

```
Out[11]:
```

| | Student | Subject | Marks |
|---|---------|-----------|-------|
| 0 | John | Physics | 78 |
| 1 | Alice | Physics | 58 |
| 2 | Rachel | Physics | 22 |
| 3 | Tom | Physics | 78 |
| 4 | John | Chemistry | 79 |

```
In [12]: # Melt df into new dataframe: df_melted
df_melted_1 = pd.melt(df, id_vars=['Student'],
```

```
value_vars= ["Physics", "Chemistry", "English", "Math"],
var_name="Subject",
value_name="Marks")
```

```
In [13]: df_melted_1
```

```
Out[13]:
```

| | Student | Subject | Marks |
|----|---------|-----------|-------|
| 0 | John | Physics | 78 |
| 1 | Alice | Physics | 58 |
| 2 | Rachel | Physics | 22 |
| 3 | Tom | Physics | 78 |
| 4 | John | Chemistry | 79 |
| 5 | Alice | Chemistry | 72 |
| 6 | Rachel | Chemistry | 61 |
| 7 | Tom | Chemistry | 89 |
| 8 | John | English | 56 |
| 9 | Alice | English | 91 |
| 10 | Rachel | English | 88 |
| 11 | Tom | English | 56 |
| 12 | John | Math | 95 |
| 13 | Alice | Math | 81 |
| 14 | Rachel | Math | 64 |
| 15 | Tom | Math | 83 |

```
In [14]: df_pivot_1 = df_melted_1.pivot(index='Student',
                                         columns='Subject', values='Marks')
```

```
In [15]: df_pivot_1
```

```
Out[15]:
```

| | Subject | Chemistry | English | Math | Physics |
|---------|---------|-----------|---------|------|---------|
| Student | | | | | |
| Alice | | 72 | 91 | 81 | 58 |
| John | | 79 | 56 | 95 | 78 |
| Rachel | | 61 | 88 | 64 | 22 |
| Tom | | 89 | 56 | 83 | 78 |

```
In [16]: import os
print(os.getcwd())
```

