Python - Analiza danych z modułem PANDAS

www.udemy.com (R)

LAB - S05-L007 - melt

1. Zaimportuj moduł pandas i numpy nadaj im standardowe aliasy. Zaimportuj też datetime, timedelta i time, możesz skorzystać z poniższych poleceń:

```
from datetime import datetime
from datetime import timedelta
import time
```

2. Do wykonania zadań z wykorzystaniem polecenia melt będziemy korzystać z danych w postaci tabeli przestawnej. Uruchom poniższy kod, który przygotuje zmienną df o odpowiedniej strukturze:

```
df = pd.read_csv('./marathon_results_2016.csv', index_col='Bib',
usecols=['Bib','40K','Half','Pace','Age','M/F','Country','State','City'])

df['40K'] = df['40K'].apply(pd.to_timedelta, errors='coerce')
df['Half'] = df['Half'].apply(pd.to_timedelta, errors='coerce')
df = df[df['40K'].notna() & df['Half'].notna()]

df['TotalSeconds'] = df['40K'].apply(lambda x: timedelta.total_seconds(x))
df['HalfSeconds'] = df['Half'].apply(lambda x: timedelta.total_seconds(x))

df = df.pivot_table(index="Age",columns="M/F",values="TotalSeconds").head()
df.head()
```

3. Usuń indeks z obiektu df

4. Zamień dane do postaci tabeli korzystając z polecenia melt definiując kolumnę **Age** jako kolumnę indeksu

5. Dodaj do poprzedniego polecenia parametr, który spowoduje, że kolumna z wartościami będzie nazwana **TotalSeconds**

6. Dodaj do poprzedniego polecenia parametr, który spowoduje, że nowo utworzona kolumna będzie miała nagłówek **Sex**

7. Podobnie jak w punkcie drugim wykonaj następujący kod, który spowoduje utworzenie nieco innego obiektu w postaci pivot table:

```
df = pd.read_csv('./marathon_results_2016.csv', index_col='Bib',
                 usecols=
['Bib','40K','Half','Pace','Age','M/F','Country','State','City'])

df['40K'] = df['40K'].apply(pd.to_timedelta, errors='coerce')
df['Half'] = df['Half'].apply(pd.to_timedelta, errors='coerce')
df = df[df['40K'].notna() & df['Half'].notna()]


df['TotalSeconds'] = df['40K'].apply(lambda x: timedelta.total_seconds(x))
df['HalfSeconds'] = df['Half'].apply(lambda x: timedelta.total_seconds(x))
```

```
 df = df.pivot_table(index="Age",columns="M/F",values=
["HalfSeconds","TotalSeconds"]).head()
 df.head()
```

8. Usuń indeks z obiektu df

9. W wesji 2.2.0 pojawił sie pewnien bug, który powodował, że melt wykonywany na kolumnach z multiindeksem nie działał. Dlatego wykonaj następne przekształcenie, które "spłaszczy multiindex do postaci zwykłego indeksu. Jeśli ten temat cię interesuje, to zajrzyj do artykułu: https://www.mobilo24.eu/python-pandas-melt-na-kolumnie-z-multi-indeksem/

```
df.columns = df.columns.to_flat_index().str.join('')
```

10. Zamień dane do postaci tabeli korzystając z polecenia melt definiując kolumnę **Age** jako kolumnę indeksu

11. Zmień poprzednie polecenie tak, aby kolumna z wartościami (aktualnie nazwana value) zmieniła nazwę na **Time**

Dane pochodzą z https://github.com/llimllib/bostonmarathon https://www.kaggle.com/rojour/boston-marathon-2016-finishers-analysis/data

Rozwiązania:

Poniżej znajdują się propozycje rozwiązań zadań. Prawdopodobnie istnieje wiele dobrych rozwiązań, dlatego jeżeli rozwiązujesz zadania samodzielnie, to najprawdopodobniej zrobisz to inaczej, może nawet lepiej :) Możesz pochwalić się swoimi rozwiązaniami w sekcji Q&A

In [1]:
```
import pandas as pd
import numpy as np
from datetime import datetime
from datetime import timedelta
import time
```

In [2]:
```
df = pd.read_csv('./marathon_results_2016.csv', index_col='Bib',
                 usecols=['Bib','40K','Half','Pace','Age','M/F',
                          'Country','State','City'])

df['40K'] = df['40K'].apply(pd.to_timedelta, errors='coerce')
df['Half'] = df['Half'].apply(pd.to_timedelta, errors='coerce')
df = df[df['40K'].notna() & df['Half'].notna()]

df['TotalSeconds'] = df['40K'].apply(lambda x: timedelta.total_seconds(x))
df['HalfSeconds'] = df['Half'].apply(lambda x: timedelta.total_seconds(x))

df = df.pivot_table(index="Age",columns="M/F",values="TotalSeconds").head()
df.head()
```

In [3]:
```
df.reset_index(inplace=True)
df.head()
```

In [4]:
```
df.melt(id_vars="Age").head(10)
```

In [5]:
```
df.melt(id_vars="Age",value_name="TotalSeconds").head()
```

In [6]:
```
df.melt(id_vars="Age",value_name="TotalSeconds",var_name="Sex").head()
```

In [7]:
```
df = pd.read_csv('./marathon_results_2016.csv', index_col='Bib',
                 usecols=['Bib','40K','Half','Pace','Age','M/F','Country',
                          'State','City'])

df['40K'] = df['40K'].apply(pd.to_timedelta, errors='coerce')
df['Half'] = df['Half'].apply(pd.to_timedelta, errors='coerce')
df = df[df['40K'].notna() & df['Half'].notna()]

df['TotalSeconds'] = df['40K'].apply(lambda x: timedelta.total_seconds(x))
df['HalfSeconds'] = df['Half'].apply(lambda x: timedelta.total_seconds(x))

df = df.pivot_table(index="Age",columns="M/F",
                    values=["HalfSeconds","TotalSeconds"]).head()
df.head()
```

In [8]:
```
df.reset_index(inplace=True)
df.head()
```

In [9]:
```
df.columns = df.columns.to_flat_index().str.join('')
```

In [10]:
```
df.melt(id_vars="Age").head()
```

In [11]:
```
df.melt(id_vars="Age", value_name='Time').head()
```

In [ ]:

In [ ]: