pandas_exercises_ANSWERS

August 5, 2019

1 Pandas exercises

[4]:
```
import pandas as pd
import numpy as np
```

1. Load the `./data/article_read.csv` file into a Dataframe. Use the headers `'my_datetime', 'event', 'country', 'user_id', 'source', 'topic'`.

[5]:
```
article_read = pd.read_csv('./data/article_read.csv', delimiter=';',␣
 ↪names=['my_datetime', 'event', 'country', 'user_id', 'source', 'topic'])
```

2. Select the user_id, the country and the topic columns for the users who are from country_2. Print the first five rows only.

[11]:
```
ar_filtered = article_read[article_read.country == 'country_2']
ar_filtered_cols = ar_filtered[['user_id','topic', 'country']]
ar_filtered_cols.head()
```

[11]:
```
       user_id   topic    country
6   2458151267  Europe  country_2
13  2458151274  Europe  country_2
17  2458151278    Asia  country_2
19  2458151280    Asia  country_2
20  2458151281    Asia  country_2
```

2. What is the most frequent source in the dataframe?

[13]:
```
article_read.groupby('source').count()[['user_id']]
```

[13]:
```
         user_id
source
AdWords      500
Reddit       949
SEO          346
```

3. For the users of country_2, what was the most frequent topic and source combination? Or in other words: which topic, from which source, brought the most views from country_2?

[15]:
```
article_read[article_read.country == 'country_2'].groupby(['source', 'topic']).
 ↪count()[['user_id']]
```

[15]:
```
                             user_id
source  topic
```

```
AdWords Africa             3
        Asia              31
        Australia          6
        Europe            46
        North America     11
        South America     14
Reddit  Africa            24
        Asia             139
        Australia         18
        Europe            29
        North America     27
        South America     26
SEO     Africa             7
        Asia               9
        Australia         10
        Europe             4
        North America     42
        South America     16
```

4. Load the `./data/blog_buy.csv` file into another Dataframe. Use the headers `'my_date_time'`, `'event'`, `'user_id'`, `'amount'`.

[7]:

```
blog_buy = pd.read_csv('./data/blog_buy.csv', delimiter=';',␣
 ↪names=['my_date_time', 'event', 'user_id', 'amount'])
```

The `article_read` dataset shows all the users who read an article on the blog, and the `blog_buy` dataset shows all the users who bought something on the very same blog between 2018-01-01 and 2018-01-07.

5. What is the average (mean) revenue between 2018-01-01 and 2018-01-07 from the users in the article_read dataframe?

[8]:

```
step_1 = article_read.merge(blog_buy, how = 'left', left_on = 'user_id',␣
 ↪right_on = 'user_id')
step_2 = step_1.amount
step_3 = step_2.fillna(0)
result = step_3.mean()
result
```

[8]:

```
1.0852367688022284
```

6. Print the top 3 countries by total revenue between 2018-01-01 and 2018-01-07.

[9]:

```
step_1 = article_read.merge(blog_buy, how = 'left', left_on = 'user_id',␣
 ↪right_on = 'user_id')
step_2 = step_1.fillna(0)
step_3 = step_2.groupby('country').sum()
step_4 = step_3.amount
step_5 = step_4.sort_values(ascending = False)
step_5.head(3)
```

[9]:

```
country
country_4    1112.0
```

```
country_5     324.0
country_2     296.0
Name: amount, dtype: float64
```

[ ]:

```
```