KEY_Practice12_Pandas-Subsetting

July 14, 2019

1 Practice: Subsetting Pandas DataFrames

For this practice, let's use the `iris` dataset:

[2]:
```
# mount Google Drive
from google.colab import drive
drive.mount('/content/gdrive')
path = '/content/gdrive/My Drive/SummerExperience-master/'
```

```
       ␣
↪---------------------------------------------------------------------------

       ModuleNotFoundError                       Traceback (most recent call␣
↪last)

       <ipython-input-2-b958c7a1dd08> in <module>
         1 # mount Google Drive
   ----> 2 from google.colab import drive
         3 drive.mount('/content/gdrive')
         4 path = '/content/gdrive/My Drive/SummerExperience-master/'


       ModuleNotFoundError: No module named 'google'
```

[3]:
```
# import pandas package
import pandas as pd
```

[4]:
```
# this is where the file is located
filename = path + 'Lessons/SampleData/iris.csv'
# load the iris dataset into a DataFrame
iris = pd.read_csv(path)
```

```
       ␣
↪---------------------------------------------------------------------------
```

```
    NameError                                 Traceback (most recent call␣
↪last)

    <ipython-input-4-a2bfd3285165> in <module>
      1 # this is where the file is located
----> 2 filename = path + 'Lessons/SampleData/iris.csv'
      3 # load the iris dataset into a DataFrame
      4 iris = pd.read_csv(path)


    NameError: name 'path' is not defined
```

Refamiliarize yourself with the dataset:

[ ]:

```python
# take a look at the beginning

iris.head()
```

Try subsetting on columns:

[ ]:

```python
# subset the species column

iris['species']
```

[ ]:

```python
# subset the sepal_length and sepal_width columns

iris[ ['sepal_length','sepal_width']]
```

Try subsetting on rows:

[5]:

```python
# subset the 2nd column

iris[iris.columns[1]]
```

```
        ␣
↪---------------------------------------------------------------------------

    NameError                                 Traceback (most recent call␣
↪last)

    <ipython-input-5-d251171e4821> in <module>
      1 # subset the 2nd column
      2
----> 3 iris[iris.columns[1]]


    NameError: name 'iris' is not defined
```

[ ]:

```
# subset the first 5 rows

iris.loc[:4]
```

[ ]:

```
# subset rows 10 through 20

iris.loc[10:20]
```

[ ]:

```
# subset rows 6, 9, and 12

iris.loc[[6,9,12]]
```

Now do both!

[ ]:

```
# subset the first 3 rows and the first 3 columns

iris.loc[:2][iris.columns[:3]]
```

[ ]:

```
# subset row 20 and the species column

iris.loc[20]['species']
```

Now let's subset using query:

[ ]:

```
# subset rows where sepal_width is greater than 4

iris.query('sepal_width > 4')
```

[ ]:

```
# subset rows where sepal_width is less than 3.5 and the species is `virginica`.

iris.query('sepal_width < 3.5 and species=="virginica"')
```

[ ]:

```
# subset rows where the pedal width is 0.3 or the species is `versicolor`.

iris.query( 'petal_width==0.3 or species=="versicolor"' )
```

[ ]:

```

```