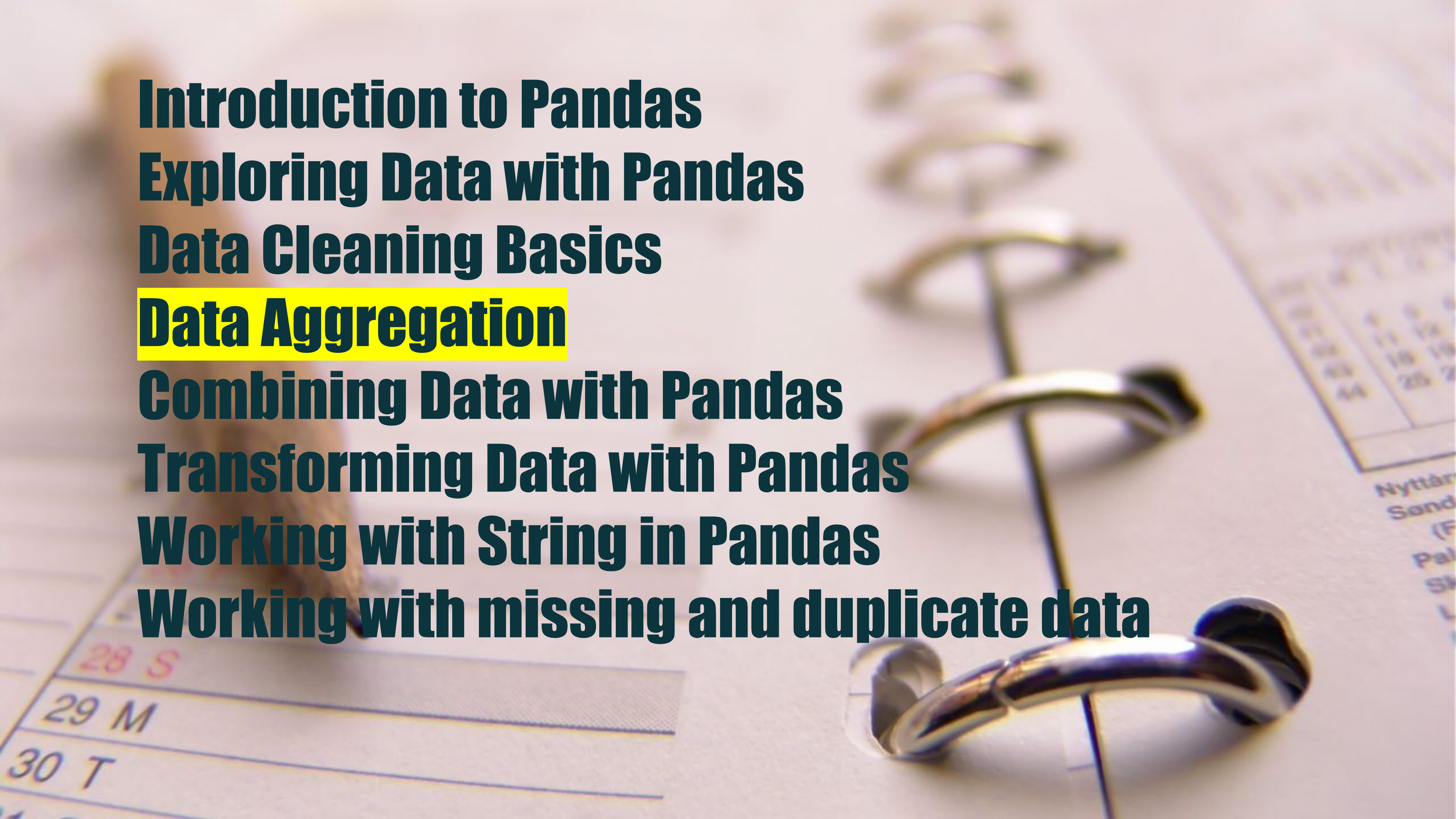
Lesson #07

Data Aggregation

March 2019



Introduction to Pandas

Exploring Data with Pandas

Data Cleaning Basics

Data Aggregation

Combining Data with Pandas

Transforming Data with Pandas

Working with String in Pandas

Working with missing and duplicate data

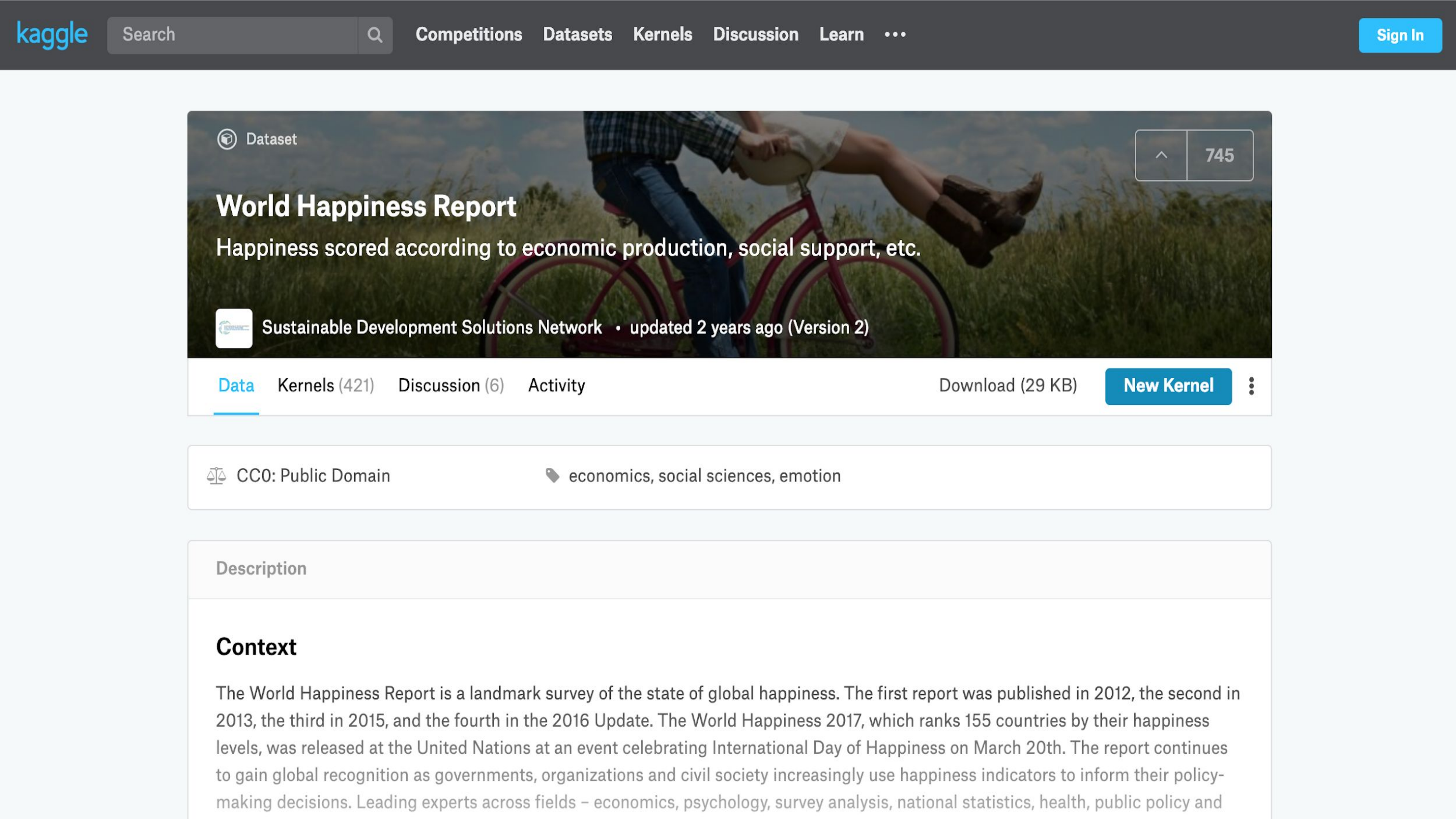
- Groupy operation
- Common aggregation methods with groupby
- Aggregation with pivot table

Update from repository

```
git clone https://github.com/ivanovitchm/datascience_one_2019_1
```

Or ….

```
git pull
```



World Happiness Report

Happiness scored according to economic production, social support, etc.

Sustainable Development Solutions Network • updated 2 years ago (Version 2)

CC0: Public Domain

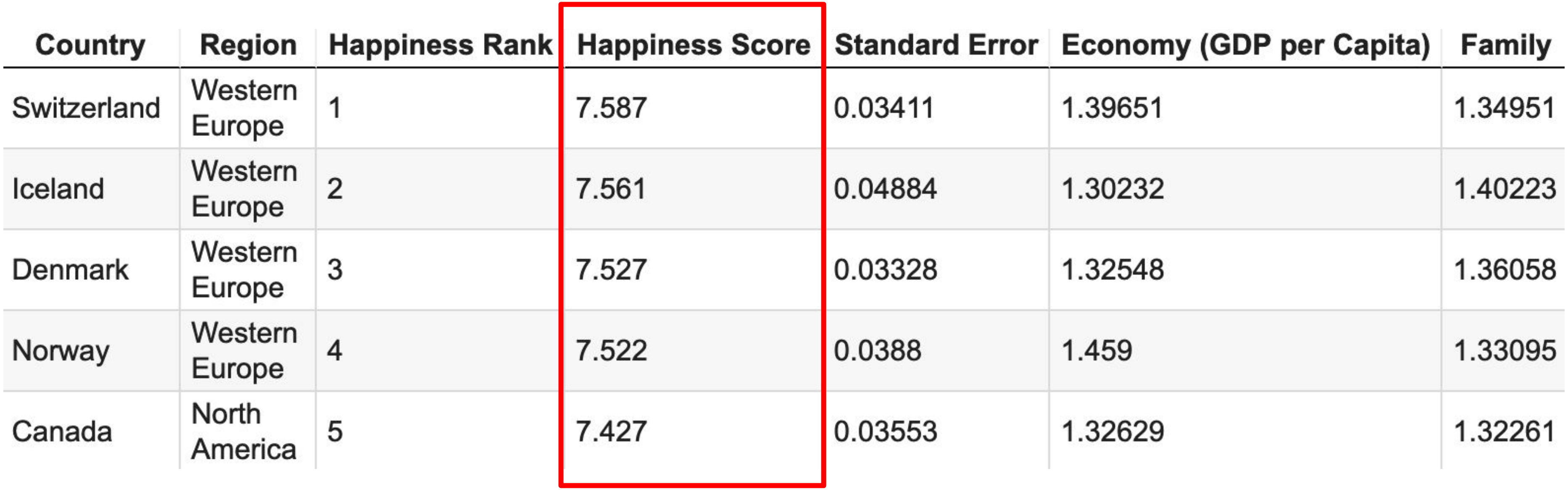
economics, social sciences, emotion

Description

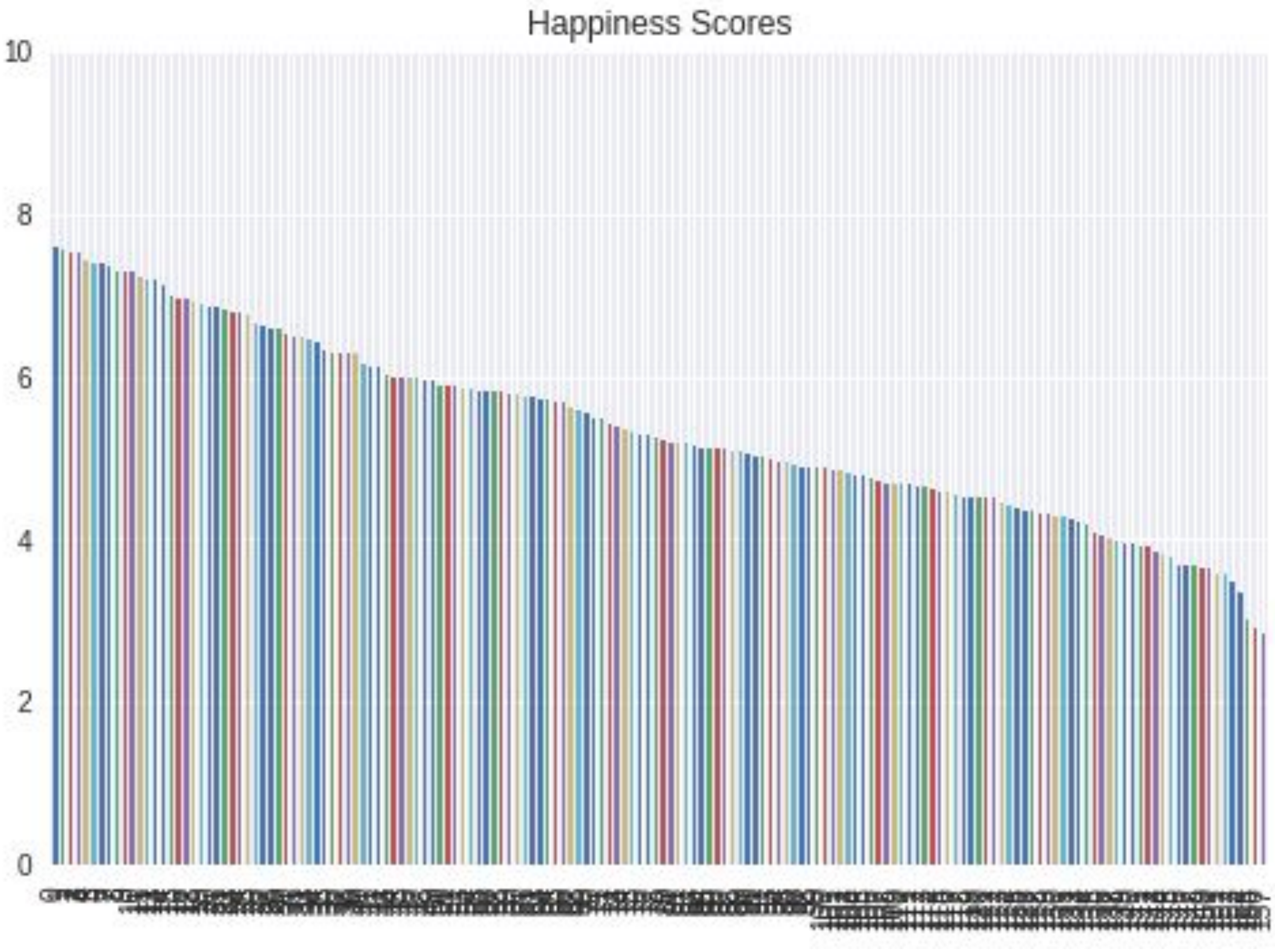
Context

The World Happiness Report is a landmark survey of the state of global happiness. The first report was published in 2012, the second in 2013, the third in 2015, and the fourth in the 2016 Update. The World Happiness 2017, which ranks 155 countries by their happiness levels, was released at the United Nations at an event celebrating International Day of Happiness on March 20th. The report continues to gain global recognition as governments, organizations and civil society increasingly use happiness indicators to inform their policy-making decisions. Leading experts across fields – economics, psychology, survey analysis, national statistics, health, public policy and

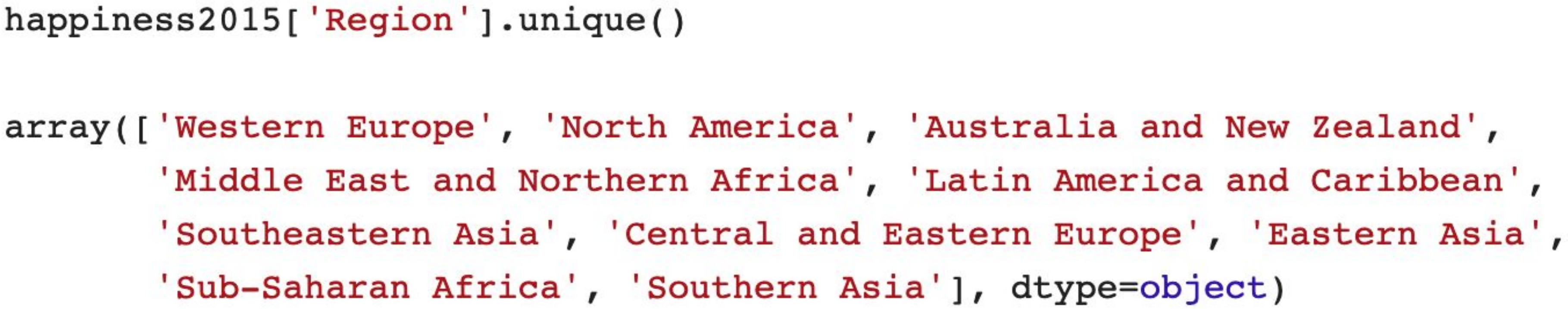
Introduction to data set



Country	Region	Happiness Rank	Happiness Score	Standard Error	Economy (GDP per Capita)	Family
Switzerland	Western Europe	1	7.587	0.03411	1.39651	1.34951
Iceland	Western Europe	2	7.561	0.04884	1.30232	1.40223
Denmark	Western Europe	3	7.527	0.03328	1.32548	1.36058
Norway	Western Europe	4	7.522	0.0388	1.459	1.33095
Canada	North America	5	7.427	0.03553	1.32629	1.32261

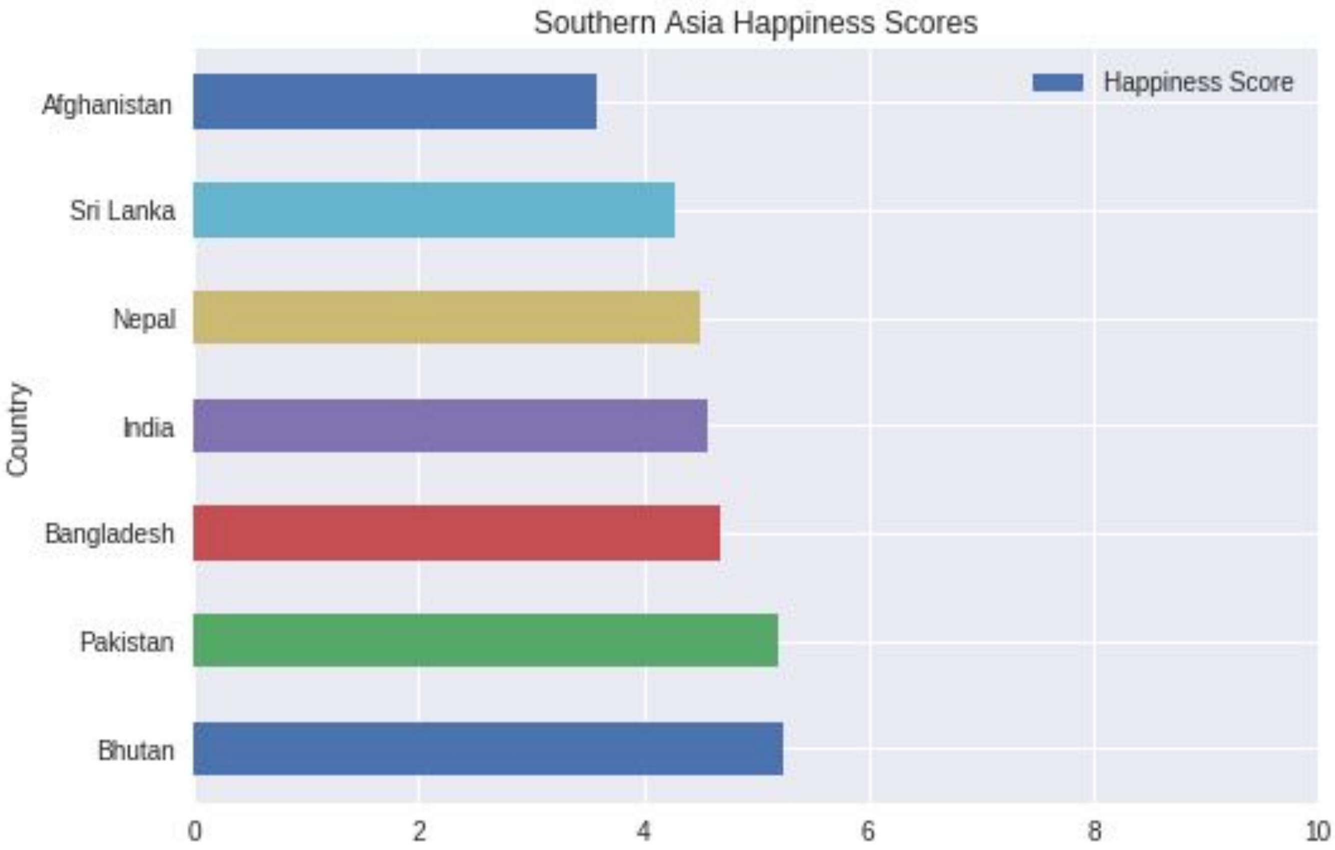

Happiness Scores
10
8
6
4
2
0

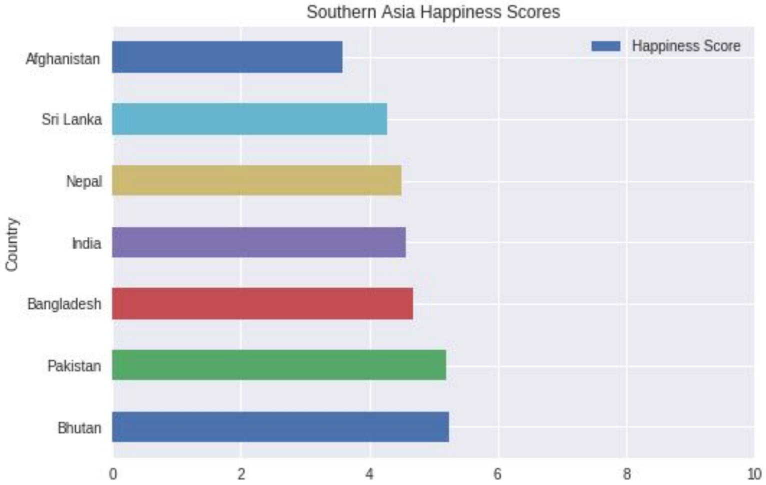
Exploring aggregation opportunities



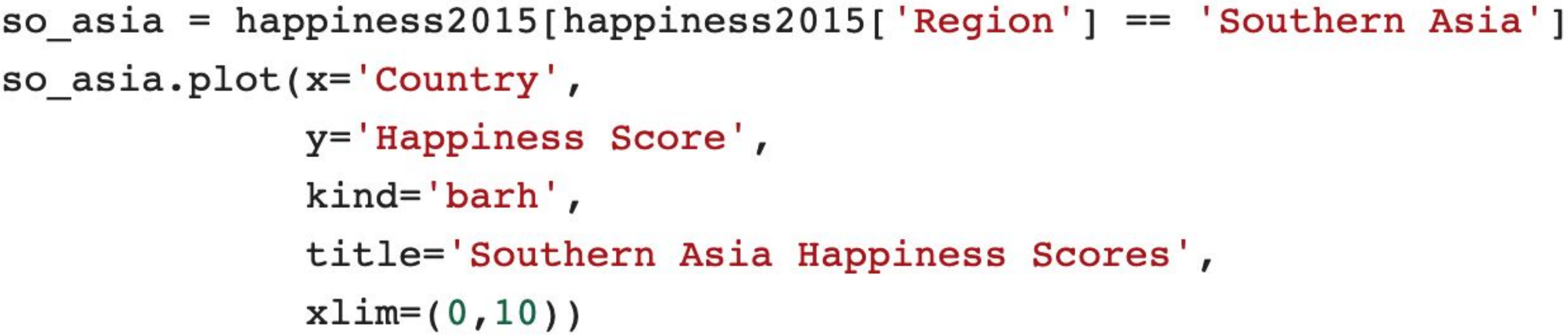
```
happiness2015['Region'].unique()

array(['Western Europe', 'North America', 'Australia and New Zealand',
       'Middle East and Northern Africa', 'Latin America and Caribbean',
       'Southeastern Asia', 'Central and Eastern Europe', 'Eastern Asia',
       'Sub-Saharan Africa', 'Southern Asia'], dtype=object)
```

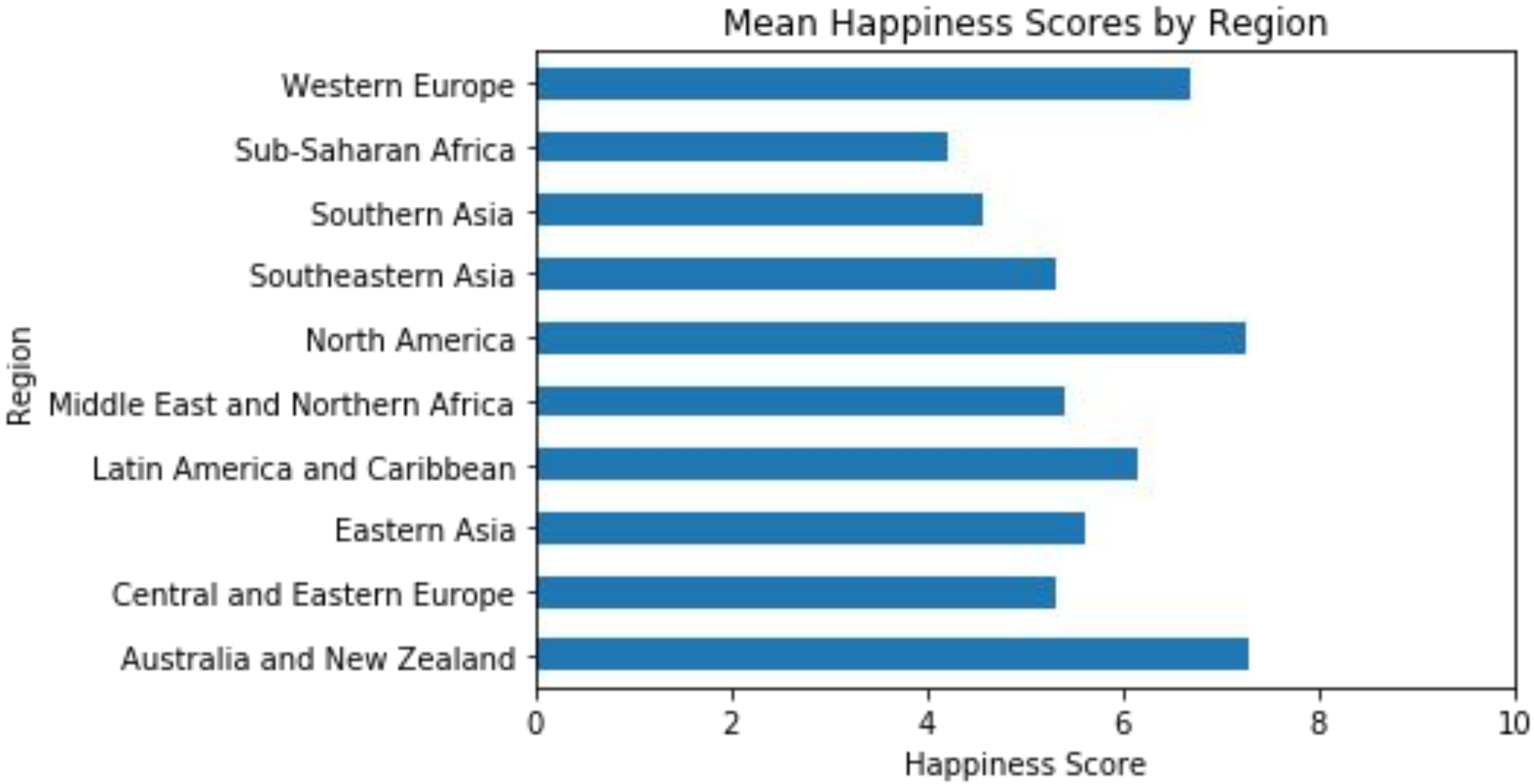

Southern Asia Happiness Scores
Happiness Score
Country
Afghanistan
Sri Lanka
Nepal
India
Bangladesh
Pakistan
Bhutan
0
2
4
6
8
10



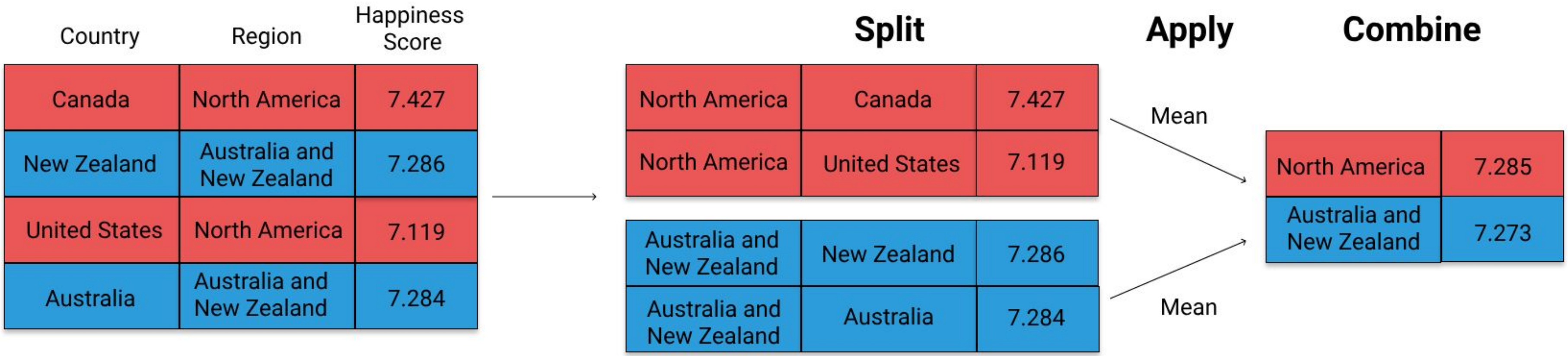
However, we wouldn't know if the Southern Asia region is representative of the entire world unless we look at the other regions



```
so_asia = happiness2015[happiness2015['Region'] == 'Southern Asia']
so_asia.plot(x='Country',
             y='Happiness Score',
             kind='barh',
             title='Southern Asia Happiness Scores',
             xlim=(0,10))
```


Mean Happiness Scores by Region
Region
Western Europe
Sub-Saharan Africa
Southern Asia
Southeastern Asia
North America
Middle East and Northern Africa
Latin America and Caribbean
Eastern Asia
Central and Eastern Europe
Australia and New Zealand
0
2
4
6
8
10
Happiness Score

The GroupBy Operation



Country	Region	Happiness Score
Canada	North America	7.427
New Zealand	Australia and New Zealand	7.286
United States	North America	7.119
Australia	Australia and New Zealand	7.284

Split

North America	Canada	7.427
North America	United States	7.119

Australia and New Zealand	New Zealand	7.286
Australia and New Zealand	Australia	7.284

Apply

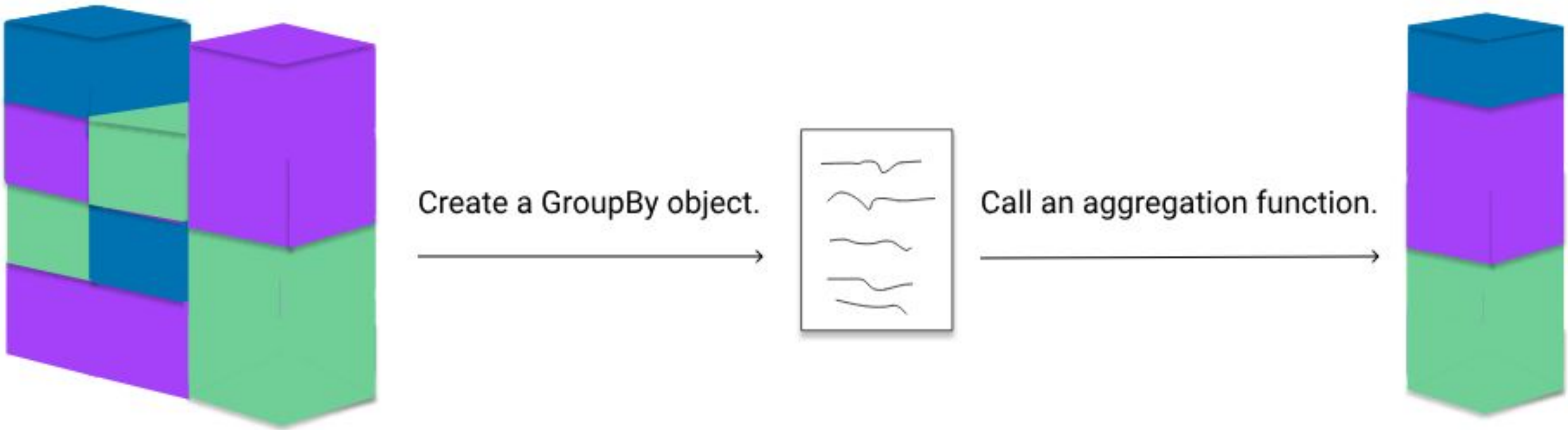
Mean

Mean

Combine

North America	7.285
Australia and New Zealand	7.273

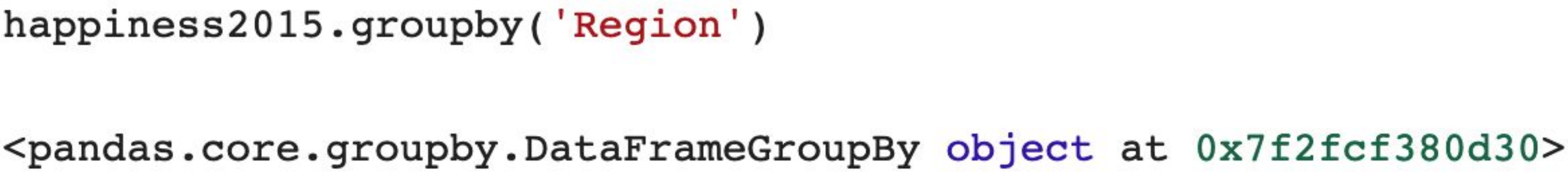
Creating GroupBy Object



Create a GroupBy object.

Call an aggregation function.

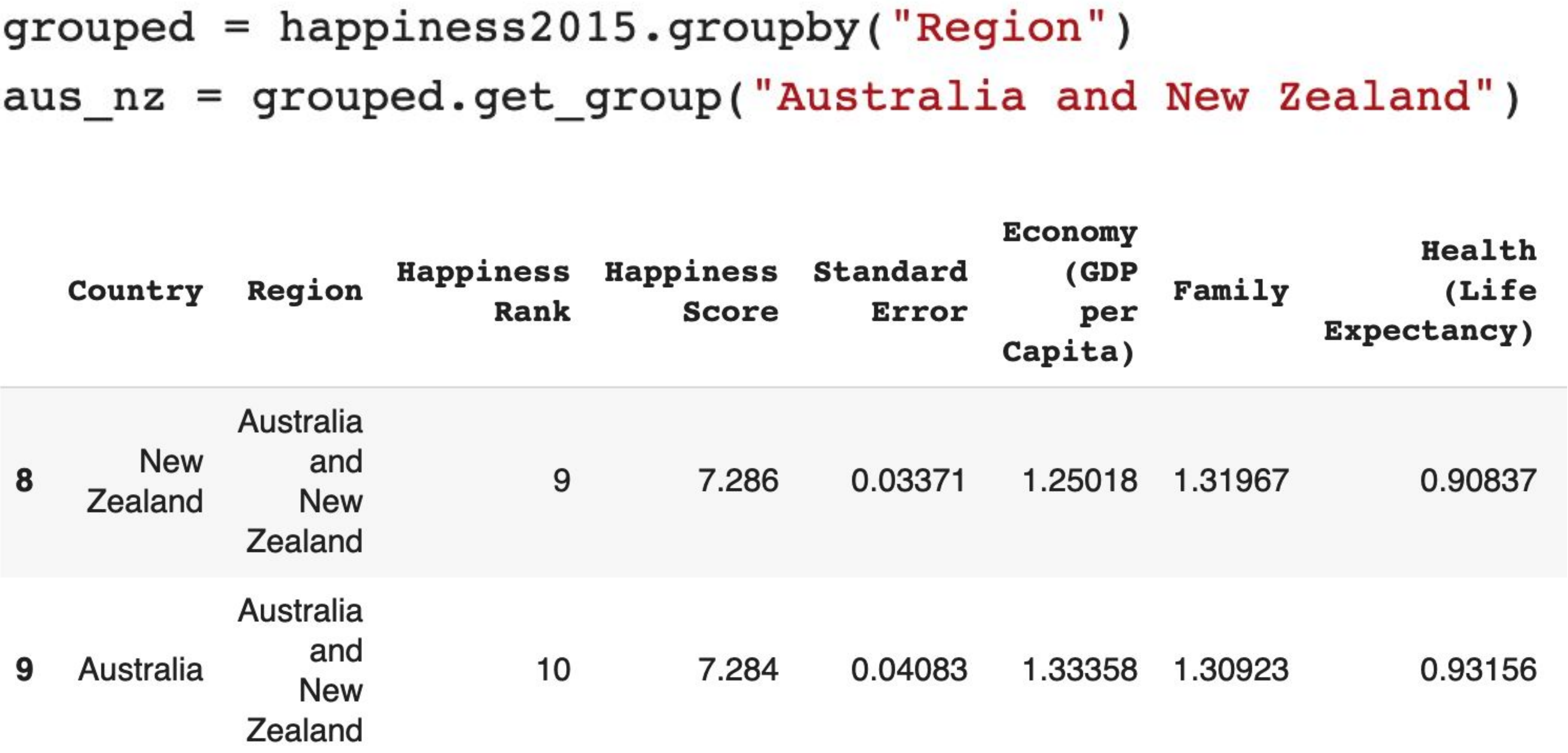
```
df.groupby('col')
```



```
happiness2015.groupby('Region')

<pandas.core.groupby.DataFrameGroupBy object at 0x7f2fcf380d30>
```

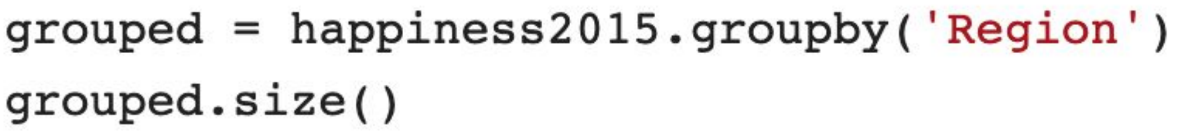
Don't be alarmed! This isn't an error. This is telling us that an object of type **GroupBy** was returned, just like we expected.



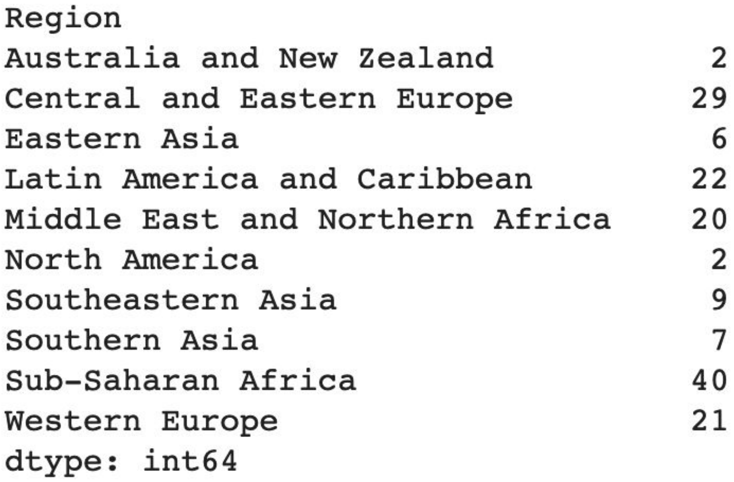
```
grouped = happiness2015.groupby("Region")
aus_nz = grouped.get_group("Australia and New Zealand")
```

	Country	Region	Happiness Rank	Happiness Score	Standard Error	Economy (GDP per Capita)	Family	Health (Life Expectancy)
8	New Zealand	Australia and New Zealand	9	7.286	0.03371	1.25018	1.31967	0.90837
9	Australia	Australia and New Zealand	10	7.284	0.04083	1.33358	1.30923	0.93156

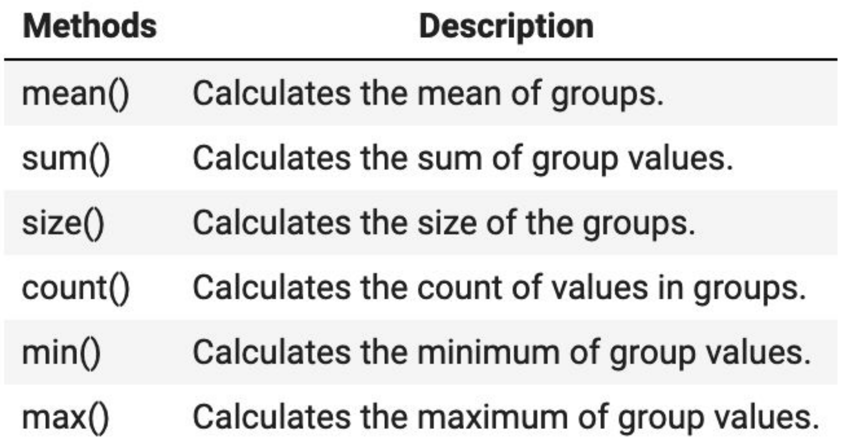
Common Aggregation Methods



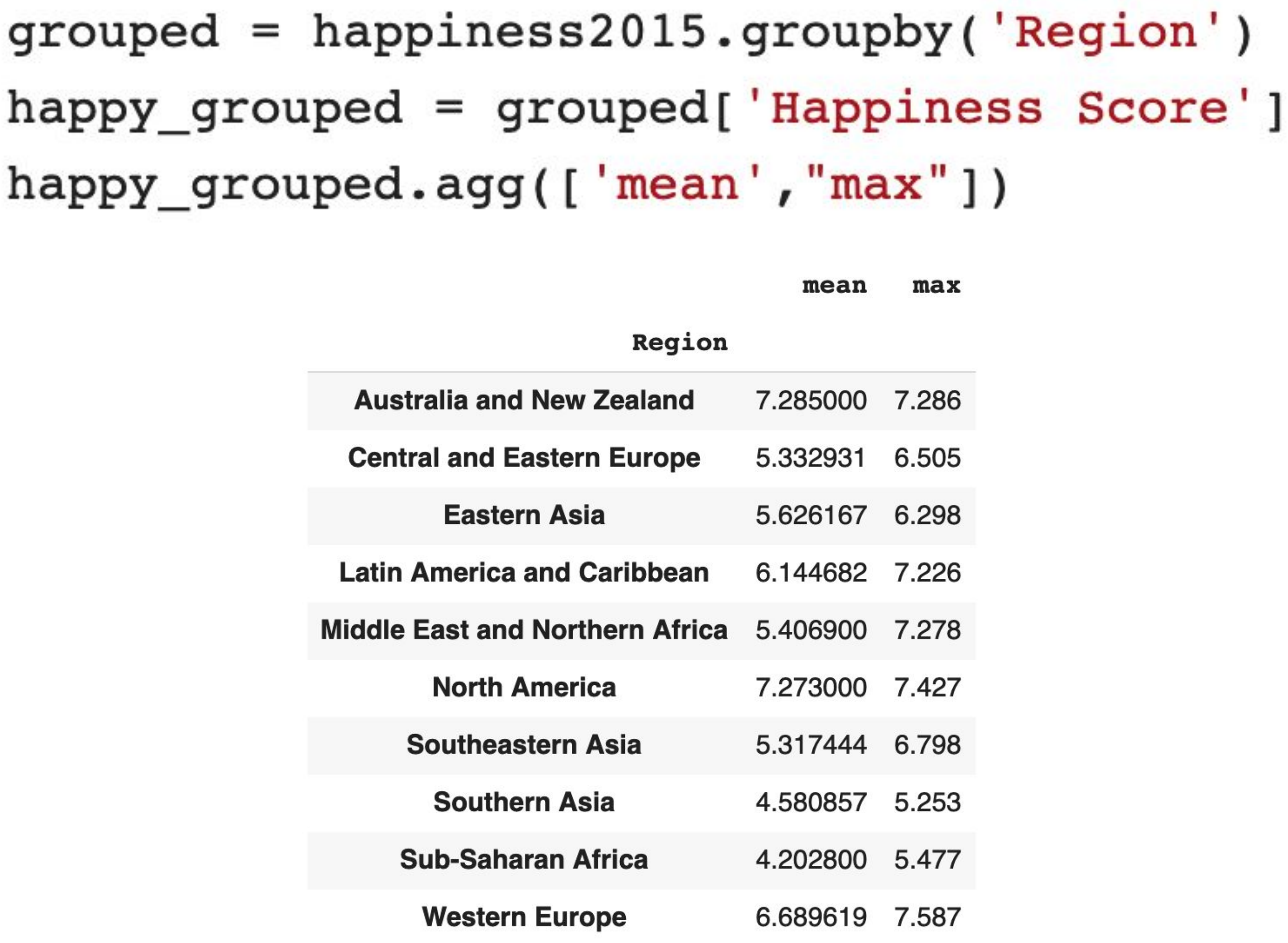
```
grouped = happiness2015.groupby('Region')
grouped.size()
```



```
Region
Australia and New Zealand              2
Central and Eastern Europe            29
Eastern Asia                           6
Latin America and Caribbean           22
Middle East and Northern Africa       20
North America                          2
Southeastern Asia                      9
Southern Asia                          7
Sub-Saharan Africa                    40
Western Europe                        21
dtype: int64
```

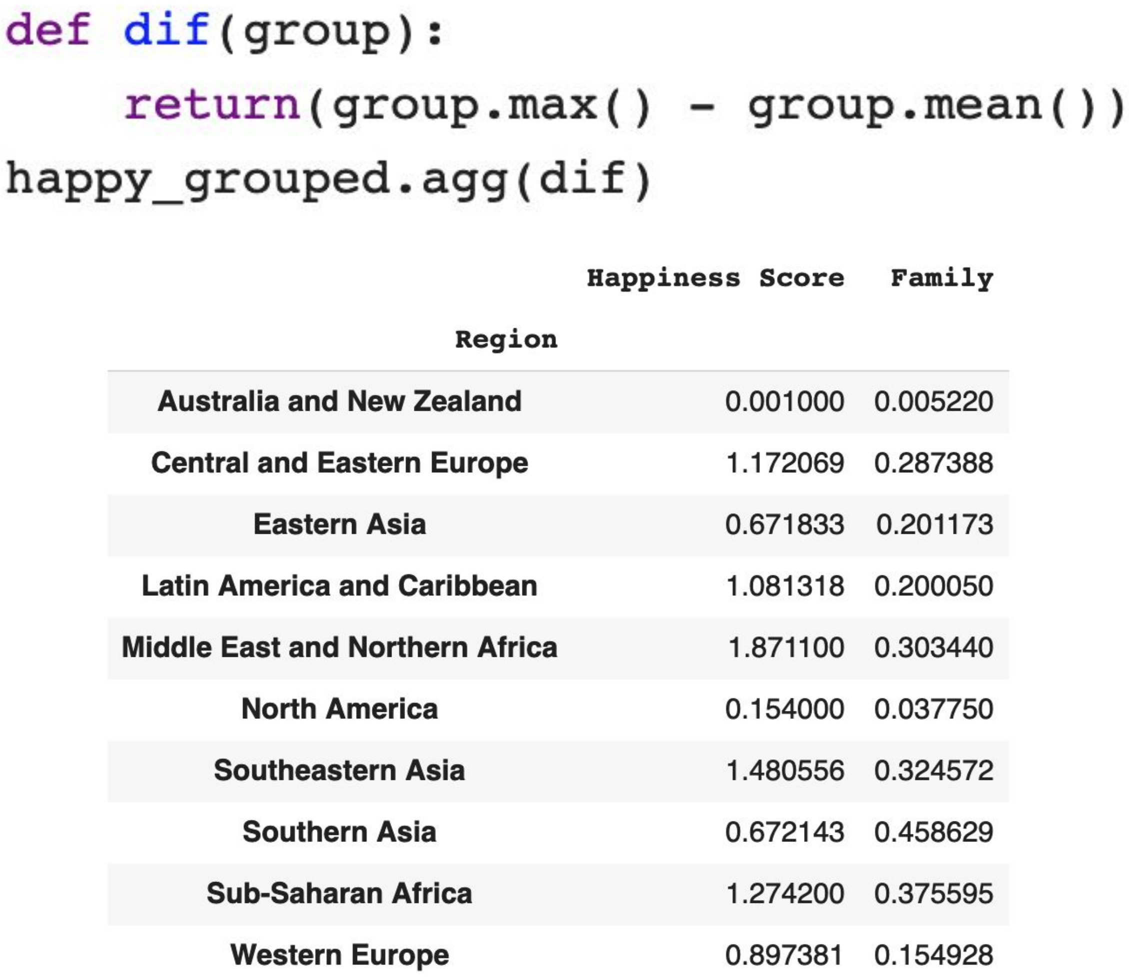


Methods	Description
mean()	Calculates the mean of groups.
sum()	Calculates the sum of group values.
size()	Calculates the size of the groups.
count()	Calculates the count of values in groups.
min()	Calculates the minimum of group values.
max()	Calculates the maximum of group values.



```
grouped = happiness2015.groupby('Region')
happy_grouped = grouped['Happiness Score']
happy_grouped.agg(['mean',"max"])
```

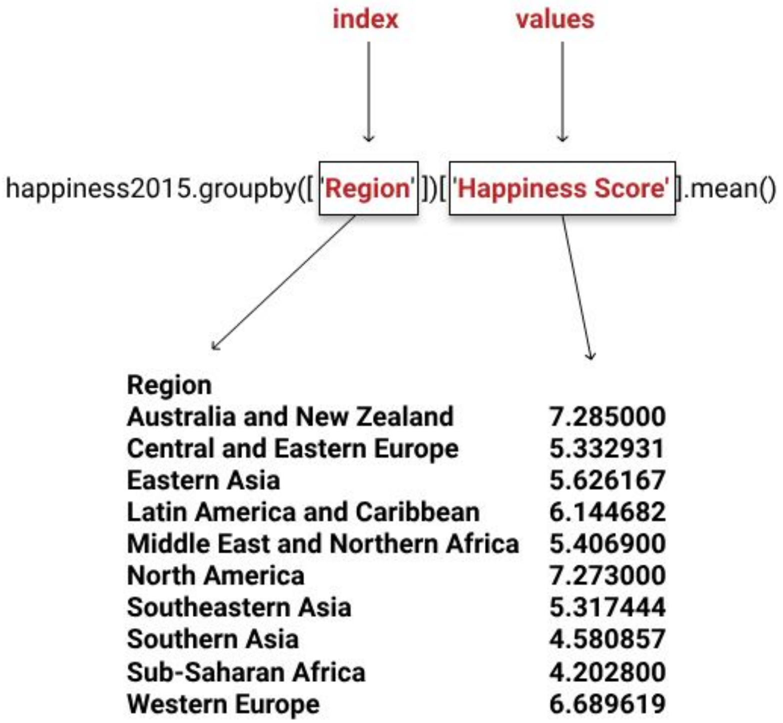
	mean	max
Region		
Australia and New Zealand	7.285000	7.286
Central and Eastern Europe	5.332931	6.505
Eastern Asia	5.626167	6.298
Latin America and Caribbean	6.144682	7.226
Middle East and Northern Africa	5.406900	7.278
North America	7.273000	7.427
Southeastern Asia	5.317444	6.798
Southern Asia	4.580857	5.253
Sub-Saharan Africa	4.202800	5.477
Western Europe	6.689619	7.587



```python
def dif(group):
    return(group.max() - group.mean())
happy_grouped.agg(dif)
```

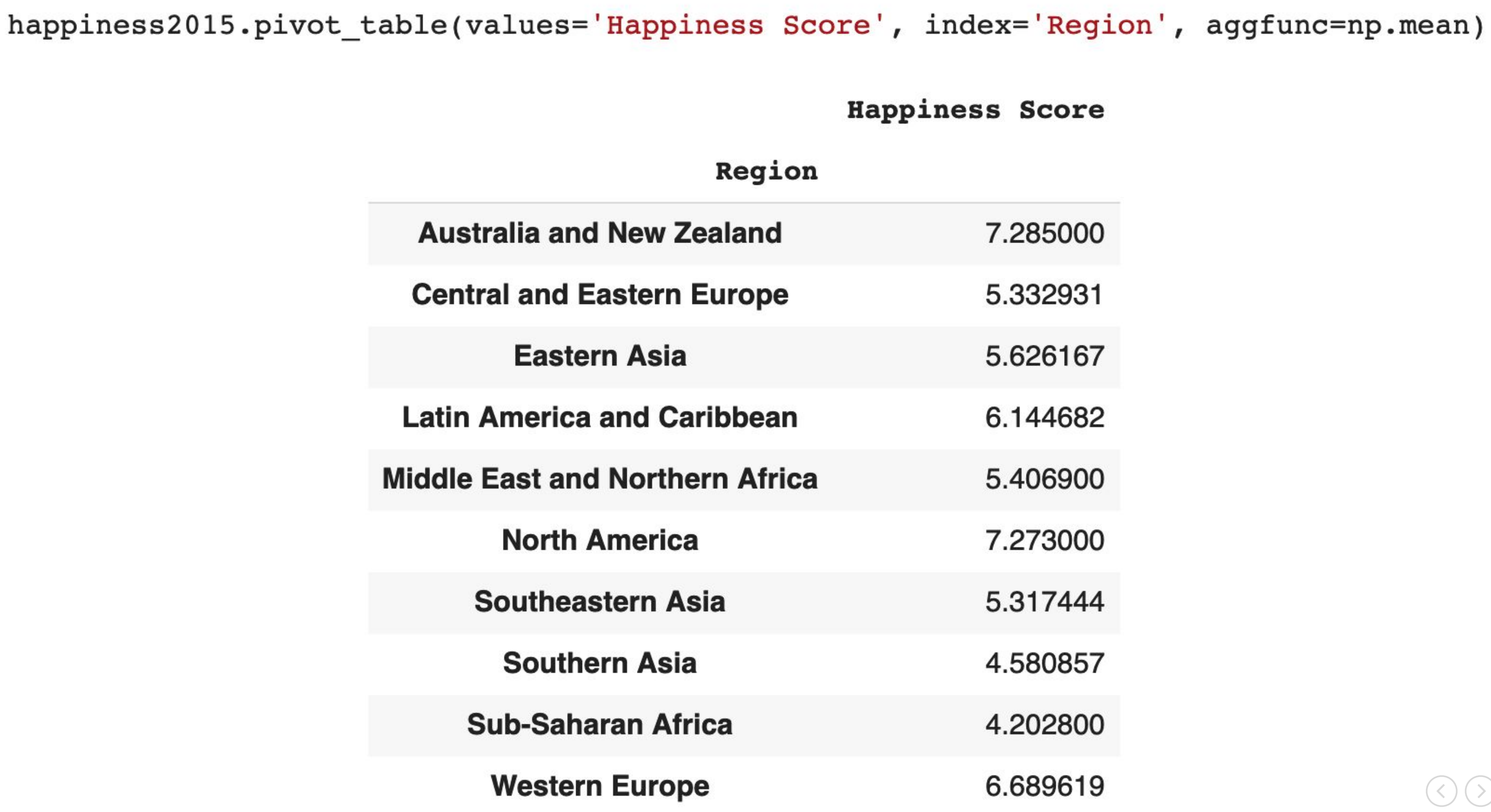
Region	Happiness Score	Family
Australia and New Zealand	0.001000	0.005220
Central and Eastern Europe	1.172069	0.287388
Eastern Asia	0.671833	0.201173
Latin America and Caribbean	1.081318	0.200050
Middle East and Northern Africa	1.871100	0.303440
North America	0.154000	0.037750
Southeastern Asia	1.480556	0.324572
Southern Asia	0.672143	0.458629
Sub-Saharan Africa	1.274200	0.375595
Western Europe	0.897381	0.154928

Aggregating with Pivot Table



```
happiness2015.groupby(['Region'])['Happiness Score'].mean()
```

```
Region
Australia and New Zealand          7.285000
Central and Eastern Europe         5.332931
Eastern Asia                       5.626167
Latin America and Caribbean        6.144682
Middle East and Northern Africa    5.406900
North America                      7.273000
Southeastern Asia                  5.317444
Southern Asia                      4.580857
Sub-Saharan Africa                 4.202800
Western Europe                     6.689619
```



```python
happiness2015.pivot_table(values='Happiness Score', index='Region', aggfunc=np.mean)
```

	Happiness Score
Region	
Australia and New Zealand	7.285000
Central and Eastern Europe	5.332931
Eastern Asia	5.626167
Latin America and Caribbean	6.144682
Middle East and Northern Africa	5.406900
North America	7.273000
Southeastern Asia	5.317444
Southern Asia	4.580857
Sub-Saharan Africa	4.202800
Western Europe	6.689619

Lesson#07 - Data Aggregation.ipynb