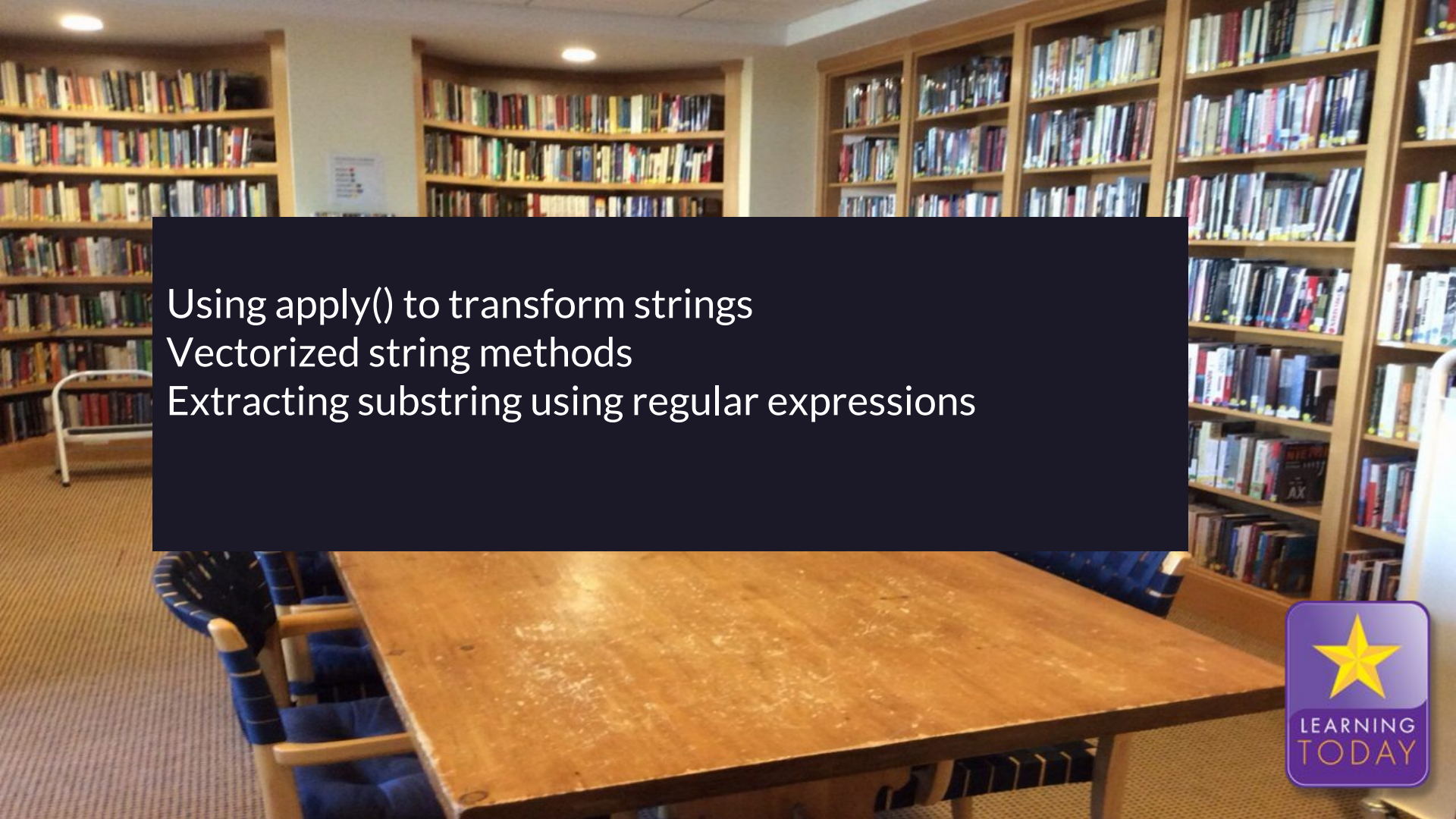
Lesson #10
Working with String in Pandas

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

Using apply() to transform strings
Vectorized string methods
Extracting substring using regular expressions



Update from repository

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

Or ....

```
git pull
```

Dataset

816

World Development Indicators

Explore country development indicators from around the world

World Bank • updated 2 years ago (Version 2)

economics, international relations

Description

The World Development Indicators from the World Bank contain over a thousand annual indicators of economic development from hundreds of countries around the world.

Here's a list of the available indicators along with a list of the available countries.

For example, this data includes the life expectancy at birth from many countries around the world:

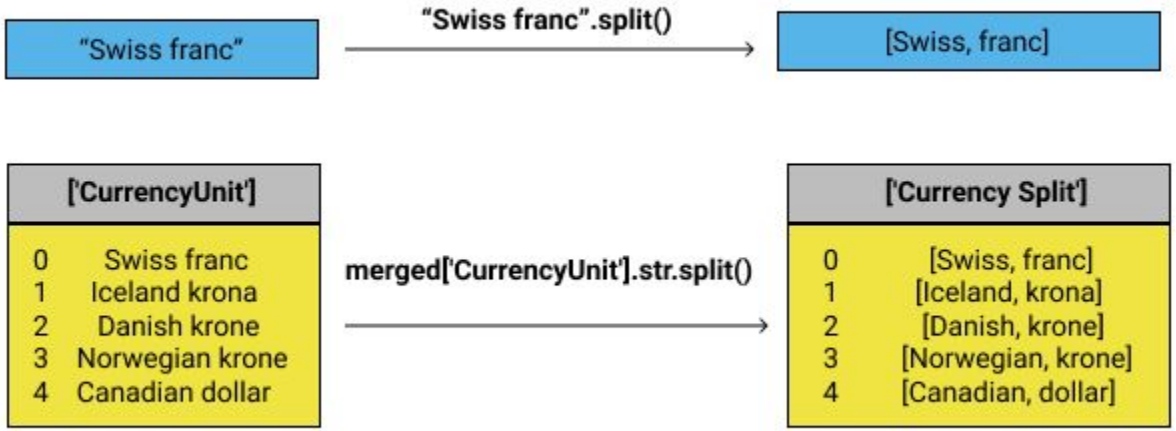
World_Happiness_2015.csv

World_dev.csv

merged

	Country	Happiness Rank	Happiness Score	CountryCode	ShortName	CurrencyUnit	IncomeGroup	SpecialNotes	IESurvey
0	Switzerland	1	7.587	CHE	Switzerland	Swiss franc	High income: OECD	NaN	Expenditure survey/budget survey (ES/BS), 2004
1	Iceland	2	7.561	ISL	Iceland	Iceland krona	High income: OECD	NaN	Integrated household survey (IHS), 2010
2	Denmark	3	7.527	DNK	Denmark	Danish krone	High income: OECD	NaN	Income tax registers (ITR), 2010
3	Norway	4	7.522	NOR	Norway	Norwegian krone	High income: OECD	NaN	Income survey (IS), 2010
4	Canada	5	7.427	CAN	Canada	Canadian dollar	High income: OECD	Fiscal year end...	Labor force survey (LFS), 2010

Vectorized String Methods Overview

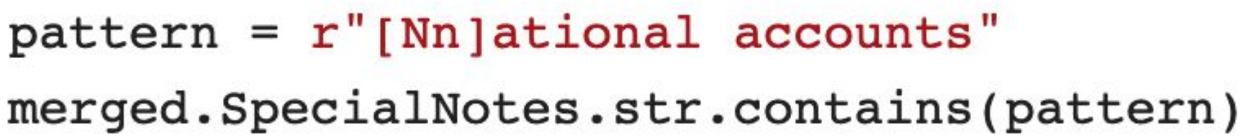


Series.str.method_name()

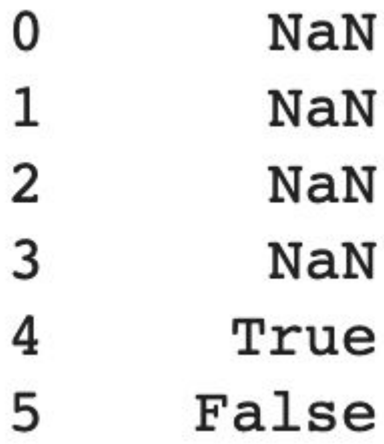
Method	Description
Series.str.split()	Splits each element in the Series.
Series.str.strip()	Strips whitespace from each string in the Series.
Series.str.lower()	Converts strings in the Series to lowercase.
Series.str.upper()	Converts strings in the Series to uppercase.
Series.str.get()	Retrieves the ith element of each element in the Series.
Series.str.replace()	Replaces a regex or string in the Series with another string.
Series.str.cat()	Concatenates strings in a Series.
Series.str.extract()	Extracts substrings from the Series matching a regex pattern.

Finding Specific Words in Strings

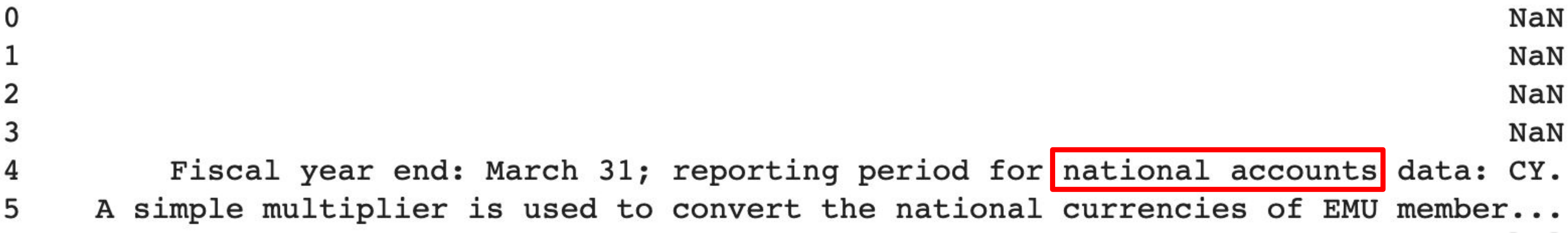
Regular Expression	Matches	Does Not Match
"ap"	"sn ap" "ap ple"	"Apple" "dog"
"123"	" 123 4" "DQ 123 4"	"4321" "1z2x3g"



```python
pattern = r"[Nn]ational accounts"
merged.SpecialNotes.str.contains(pattern)
```

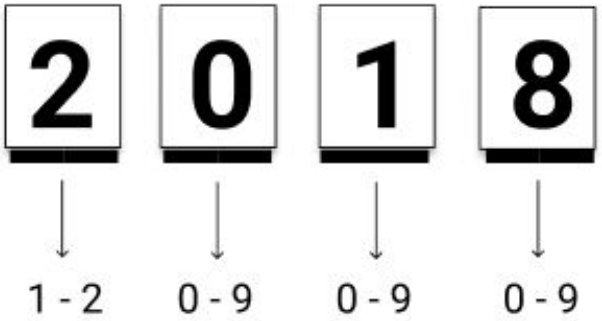


```
0      NaN
1      NaN
2      NaN
3      NaN
4     True
5    False
```



```
0                                                                                     NaN
1                                                                                     NaN
2                                                                                     NaN
3                                                                                     NaN
4      Fiscal year end: March 31; reporting period for national accounts data: CY.
5   A simple multiplier is used to convert the national currencies of EMU member...
```

```
153    2006
154     NaN
155     NaN
156     NaN
157    2013
```




```
pattern = r"([1-2][0-9]{3})"
merged.SpecialNotes.str.extract(pattern).tail()
```

Extracting All Matches of a Pattern from a Series

```
pattern = r"(?P<Years>[1-2][0-9]{3})"
merged['SpecialNotes'].str.extractall(pattern).head()
```

		Years
Country	**match**	
Finland	**0**	1999
	1	1999
Netherlands	**0**	1999
	1	2037
	2	1999

Lesson #10
Working With Strings in Pandas