BTS | Barcelona Technology School

# Data Science Foundations

**Master in Big Data Solutions 2019-2020**

Víctor Pajuelo

victor.pajuelo@bts.tech

# First things first

```
$ git clone https://github.com/vfp1/bts-mbds-data-science-foundations-2019.git
# In case you didn't do it yet

$ git fetch  # There are changes!

$ git checkout master  # Just in case

$ git merge --ff-only origin/master  # If in error, you probably made some commits to master
```

# Today's Objective

- Understand the basic abstractions of pandas
  - Series, DataFrames and Indexes
- Explore techniques to read standard text files
- Learn how to do basic data exploration
- Do basic analytics with pandas

- Why is this useful for a digital project?
  - Understanding data manipulation is crucial for any data science project

# Contents

1. Basic data operations on relational data (Pandas Introduction)
   - Recap of input/output and on-disk formats
   - Cleaning noisy data, normalizing
   - Filtering rows and columns
   - Joining data from multiple sources
   - Split-apply-combine workflows: groupby

# Recap

# Recap

- Scalar
  - Index
    - Series
      - DataFrame

- Description of DataFrame
- Sorting techniques
- Selection (.at, .iat, .loc, .iloc)
- Missing Data
- Operations (Stats, apply)
- Plotting
- IO (Input/Output)

# Slice Notation

```
import numpy as np
a = np.asarray([0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10])
```

```
a[0:11]
```

```
array([ 0,  1,  2,  3,  4,  5,  6,  7,  8,  9, 10])
```

```
a[5:]
```

```
array([ 5,  6,  7,  8,  9, 10])
```

```
a[:5]
```

```
array([0, 1, 2, 3, 4])
```

```
a[:]
```

```
array([ 0,  1,  2,  3,  4,  5,  6,  7,  8,  9, 10])
```

# Slice Notation

```
a[:]
```

```
array([ 0,  1,  2,  3,  4,  5,  6,  7,  8,  9, 10])
```

```
a[-1]
```

```
10
```

```
a[-2:]
```

```
array([ 9, 10])
```

```
a[:-2]
```

```
array([0, 1, 2, 3, 4, 5, 6, 7, 8])
```

```
a[::-1]
```

```
array([10,  9,  8,  7,  6,  5,  4,  3,  2,  1,  0])
```

```
a[1::-1]
```

```
array([1, 0])
```

```
a[-3::-1]
```

```
array([8, 7, 6, 5, 4, 3, 2, 1, 0])
```

# Pandas introduction continued

# Pandas continued

- Setting options in Pandas
- Categoricals
- Boolean indexing
- Grouping (GroupBy)

First part of class

- Time series
- Merge
- Reshaping
- Cleaning data
- Advanced IO

Second part of class

- Practice!

Third part of class

# Pandas

**Let’s go to the code!**

Go to the notebook called 02_Pandas_continued.ipynb

You have it in git and in the files section in teams


BARCELONA
Barcelona Technology School