

Data Analysis in Python: pandas in Practice

Jan Smitka

jan.smitka@lynt.cz

@jansmitka

Lynt services s.r.o.



<https://lynt.cz>



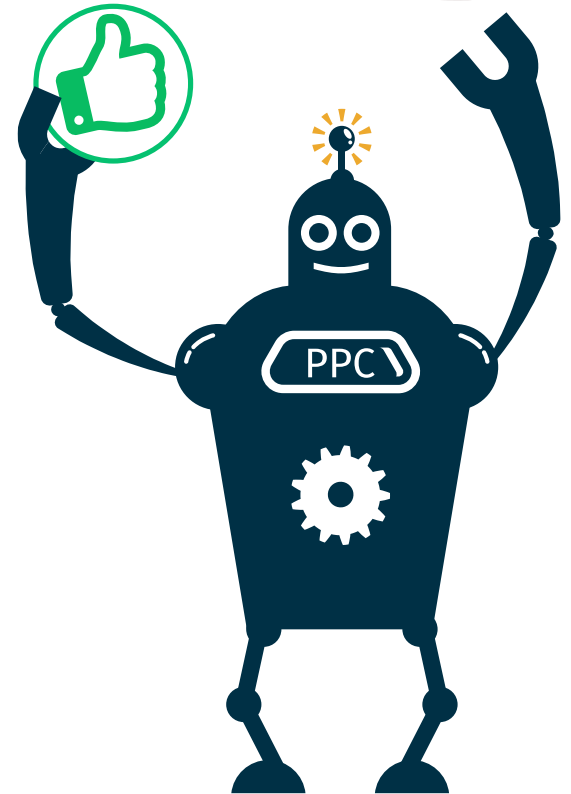
[@jansmitka](https://twitter.com/jansmitka)

Slides: <https://u.lynt.cz/pandas>

About Me



Lynt
SERVICES





Code

Issues

Pull requests

Projects

Wiki

Insights

Settings

Slides and code examples for a talk about pandas, which was presented at Pilsen Pyvo on 9th May, 2018.

Edit

Add topics

1 commit

1 branch

0 releases

1 contributor

CC-BY-4.0

Branch: master

New pull request

Create new file

Upload files

Find file

Clone or download



jontika Initial version of the talk data, notebook and slides.

Latest commit 418271c 2 minutes ago



data Initial version of the talk data, notebook and slides.

2 minutes ago



generator Initial version of the talk data, notebook and slides.

2 minutes ago



notebooks Initial version of the talk data, notebook and slides.

2 minutes ago



output Initial version of the talk data, notebook and slides.

2 minutes ago



slides Initial version of the talk data, notebook and slides.

2 minutes ago



.gitignore Initial version of the talk data, notebook and slides.

2 minutes ago



LICENSE.md Initial version of the talk data, notebook and slides.

2 minutes ago



README.md Initial version of the talk data, notebook and slides.

2 minutes ago



requirements.txt Initial version of the talk data, notebook and slides.

2 minutes ago



README.md Initial version of the talk data, notebook and slides.

2 minutes ago



README.md Initial version of the talk data, notebook and slides.

2 minutes ago

Data Analysis in Python: pandas in practice



MOTIVATION



Why not just use Excel, Google Sheets, etc.?















































[illegible]

Quite a list!













[illegible]

Still going...





[illegible]

Are we there yet?

















That was so long!



Automation is a necessity



Why not just use pure Python?

Better performance

No need to reinvent the wheel

Say hello to pandas!

- open-source (BSD), high-performance data analysis library
- supported by the NumFOCUS non-profit organization under their PyData program
 - IPython, matplotlib, NumPy, Anaconda and many other tools
- based on NumPy – many functions can be combined



INTRODUCTION

Installation

```
pip install pandas
```

Recommended libraries

- **bottleneck**: accelerates several operations related to NaN handling
- **xlrd**: enables reading of XLS and XLSX files
- **XlsxWriter**: enables writing of XLSX files
- **SQLAlchemy**: reading/writing from/to SQL databases, requires a database driver, such as pymysql or psycopg2

Data structures: Series

`pandas.Series`



homogeneous 1-dimensional array,
each element has the same dtype

Data structures: DataFrame

`pandas.DataFrame`

Index	ID	Name	Age	Job

Collection of named Series with index.

Data types (dtype)

fixed-size numerics:

- bool
- int32, int64
- float64
- datetime64[ns]
- timedelta[ns]

arbitrary data:

- category (like enum)
- object

integers have limited size!



A REAL WORLD EXAMPLE

Source of Fake Data

e-sportshop.cz
u nás dostanete vždy něco navíc...

Přihlásit se | Registrace | Kontakt | Čeština ▾

Potřebujete poradit?
Jsme tu pro Vás: **603 452 982**

Vyhledávání Hledat 🔍

Nákupní košík
Košík je prázdný.

Nejprodávanejší

- Tenis
- Fotbal
- Fitness
- Hokej
- Badminton
- Florbal
- Tréninkové pomůcky
- Vybavení sportoviště
- Sportovní radovánky

CELÝ SORTIMENT

VŠECHNY SPORTY

ZVÝHODNĚNÉ PRODUKTY

VELKÝ MEZISEZÓNÍ VÝPRODEJ

Dárek zdarma
Ke každé objednávce nad 1000 Kč.

Výhodné doručení
Poštovné od 75 Kč, u objednávek nad 3000 Kč zdarma.

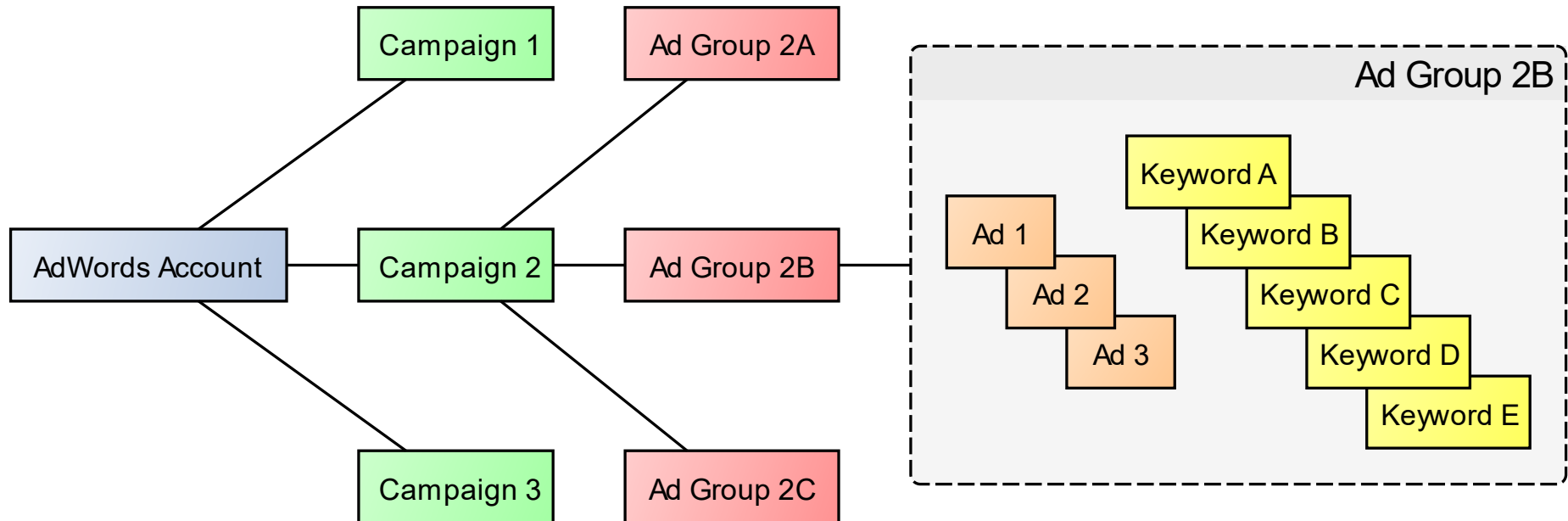
Nabídky týdne
Mimořádné slevy každý týden.

Ověřeno zákazníky
99% zákazníků nás doporučuje na [Heureka.cz](https://www.heureka.cz).

Tipy a výhodné nabídky

Professional zástěna na kurty, AKCE Akce	Yoga Foam Roller LS3768C Akce	balanční míč BBT Akce	GP1600 boxovací trenažér Akce
---	---	---------------------------------	--

AdWords Account Structure



File Edit View Run Kernel Tabs Settings Help											
Files	Code										
	Out[21]:	CampaignId	CampaignName	AdGroupId	AdGroupName	Date	Impressions	Clicks	Cost	Conversions	Converted
Running	0	10	Sortiment	100002	BANDAŽE & LÉKÁRNAČKA - Bandaje	2018-01-01	8	0	0.00	0	
	1	10	Sortiment	100002	BANDAŽE & LÉKÁRNAČKA - Bandaje	2018-01-02	10	0	0.00	0	
Commands	2	10	Sortiment	100002	BANDAŽE & LÉKÁRNAČKA - Bandaje	2018-01-03	23	2	7.80	0	
	3	10	Sortiment	100002	BANDAŽE & LÉKÁRNAČKA - Bandaje	2018-01-04	10	0	0.00	0	
	4	10	Sortiment	100002	BANDAŽE & LÉKÁRNAČKA - Bandaje	2018-01-05	8	0	0.00	0	
	5	10	Sortiment	100002	BANDAŽE & LÉKÁRNAČKA - Bandaje	2018-01-06	6	0	0.00	0	
Cell Tools	6	10	Sortiment	100002	BANDAŽE & LÉKÁRNAČKA - Bandaje	2018-01-07	4	0	0.00	0	
	7	10	Sortiment	100002	BANDAŽE & LÉKÁRNAČKA - Bandaje	2018-01-08	8	0	0.00	0	
Tabs	8	10	Sortiment	100002	BANDAŽE & LÉKÁRNAČKA - Bandaje	2018-01-09	8	0	0.00	0	
	9	10	Sortiment	100002	BANDAŽE & LÉKÁRNAČKA - Bandaje	2018-01-10	11	0	0.00	0	
Open the Jupyter Notebook											
	10	10	Sortiment	100002	BANDAŽE & LÉKÁRNAČKA - Bandaje	2018-01-11	8	0	0.00	0	
	11	10	Sortiment	100002	BANDAŽE & LÉKÁRNAČKA - Bandaje	2018-01-12	5	0	0.00	0	
	12	10	Sortiment	100002	BANDAŽE & LÉKÁRNAČKA - Bandaje	2018-01-13	5	0	0.00	0	
	13	10	Sortiment	100002	BANDAŽE & LÉKÁRNAČKA - Bandaje	2018-01-14	3	0	0.00	0	
	14	10	Sortiment	100002	BANDAŽE & LÉKÁRNAČKA - Bandaje	2018-01-15	15	1	2.85	0	
	15	10	Sortiment	100002	BANDAŽE & LÉKÁRNAČKA - Bandaje	2018-01-16	20	1	8.04	0	
	16	10	Sortiment	100002	BANDAŽE & LÉKÁRNAČKA - Bandaje	2018-01-17	30	4	15.30	0	
	17	10	Sortiment	100002	BANDAŽE & LÉKÁRNAČKA - Bandaje	2018-01-18	13	0	0.00	0	
	18	10	Sortiment	100002	BANDAŽE & LÉKÁRNAČKA - Bandaje	2018-01-19	13	1	4.45	0	
	19	10	Sortiment	100002	BANDAŽE & LÉKÁRNAČKA - Bandaje	2018-01-20	10	0	0.00	0	
	20	10	Sortiment	100002	BANDAŽE & LÉKÁRNAČKA - Bandaje	2018-01-21	6	0	0.00	0	
	21	10	Sortiment	100002	BANDAŽE & LÉKÁRNAČKA - Bandaje	2018-01-22	23	1	3.88	0	
	22	10	Sortiment	100002	BANDAŽE & LÉKÁRNAČKA - Bandaje	2018-01-23	35	2	7.58	0	

Common Metrics in AdWords

CTR (Click-Through Rate):

$$CTR = \frac{Clicks}{Impressions}$$

CPC (Cost Per Click):

$$CPC = \frac{Cost}{Clicks}$$

Average Conversion Value:

$$AvgConversionValue = \frac{ConversionsValue}{Conversions}$$

Files	File Edit View Run Kernel Tabs Settings Help										
	<div> </div>										
Running	Out[21]:										
		CampaignId	CampaignName	AdGroupId	AdGroupName	Date	Impressions	Clicks	Cost	Conversions	Converted
Commands	0	10	Sortiment	100002	BANDAŽE & LÉKÁRNAČKA - Bandaje	2018-01-01	8	0	0.00	0	
	1	10	Sortiment	100002	BANDAŽE & LÉKÁRNAČKA - Bandaje	2018-01-02	10	0	0.00	0	
	2	10	Sortiment	100002	BANDAŽE & LÉKÁRNAČKA - Bandaje	2018-01-03	23	2	7.80	0	
	3	10	Sortiment	100002	BANDAŽE & LÉKÁRNAČKA - Bandaje	2018-01-04	10	0	0.00	0	
	4	10	Sortiment	100002	BANDAŽE & LÉKÁRNAČKA - Bandaje	2018-01-05	8	0	0.00	0	
	5	10	Sortiment	100002	BANDAŽE & LÉKÁRNAČKA - Bandaje	2018-01-06	6	0	0.00	0	
Cell Tools	6	10	Sortiment	100002	BANDAŽE & LÉKÁRNAČKA - Bandaje	2018-01-07	4	0	0.00	0	
	7	10	Sortiment	100002	BANDAŽE & LÉKÁRNAČKA - Bandaje	2018-01-08	8	0	0.00	0	
	8	10	Sortiment	100002	BANDAŽE & LÉKÁRNAČKA - Bandaje	2018-01-09	5	0	0.00	0	
	9	10	Sortiment	100002	BANDAŽE & LÉKÁRNAČKA - Bandaje	2018-01-10	11	0	0.00	0	
Tabs	10	10	Sortiment	100002	BANDAŽE & LÉKÁRNAČKA - Bandaje	2018-01-11	5	0	0.00	0	
	11	10	Sortiment	100002	BANDAŽE & LÉKÁRNAČKA - Bandaje	2018-01-12	5	0	0.00	0	
	12	10	Sortiment	100002	BANDAŽE & LÉKÁRNAČKA - Bandaje	2018-01-13	5	0	0.00	0	
	13	10	Sortiment	100002	BANDAŽE & LÉKÁRNAČKA - Bandaje	2018-01-14	3	0	0.00	0	
	14	10	Sortiment	100002	BANDAŽE & LÉKÁRNAČKA - Bandaje	2018-01-15	15	1	2.85	0	
	15	10	Sortiment	100002	BANDAŽE & LÉKÁRNAČKA - Bandaje	2018-01-16	20	1	8.04	0	
	16	10	Sortiment	100002	BANDAŽE & LÉKÁRNAČKA - Bandaje	2018-01-17	30	4	15.30	0	
	17	10	Sortiment	100002	BANDAŽE & LÉKÁRNAČKA - Bandaje	2018-01-18	13	0	0.00	0	
	18	10	Sortiment	100002	BANDAŽE & LÉKÁRNAČKA - Bandaje	2018-01-19	13	1	4.45	0	
	19	10	Sortiment	100002	BANDAŽE & LÉKÁRNAČKA - Bandaje	2018-01-20	10	0	0.00	0	
	20	10	Sortiment	100002	BANDAŽE & LÉKÁRNAČKA - Bandaje	2018-01-21	6	0	0.00	0	
	21	10	Sortiment	100002	BANDAŽE & LÉKÁRNAČKA - Bandaje	2018-01-22	23	1	5.88	0	
	22	10	Sortiment	100002	BANDAŽE & LÉKÁRNAČKA - Bandaje	2018-01-23	35	2	7.58	0	

Return to Jupyter Notebook, section Computations

SQL-like joins: left join

left

key	name
A	Arthur Dent
B	Ford Prefect
C	Tricia McMillan

right

key	drink
C	Wine
D	That Old Janx Spirit
A	Tea

merge(how='left')

key	name	drink
A	Arthur Dent	Tea
B	Ford Prefect	NaN
C	Tricia McMillan	Wine

SQL-like joins: right join

left

key	name
A	Arthur Dent
B	Ford Prefect
C	Tricia McMillan

right

key	drink
C	Wine
D	That Old Janx Spirit
A	Tea

merge(how='right')

key	name	drink
C	Tricia McMillan	Wine
D	NaN	That Old Janx Spirit
A	Arthur Dent	Tea

SQL-like joins: inner join

left

key	name
A	Arthur Dent
B	Ford Prefect
C	Tricia McMillan

right

key	drink
C	Wine
D	That Old Janx Spirit
A	Tea

merge(how='inner')

key	name	drink
A	Arthur Dent	Tea
C	Tricia McMillan	Wine

SQL-like joins: outer join

left

key	name
A	Arthur Dent
B	Ford Prefect
C	Tricia McMillan

right

key	drink
C	Wine
D	That Old Janx Spirit
A	Tea

merge(how='outer')

key	name	drink
A	Arthur Dent	Tea
B	Ford Prefect	NaN
D	NaN	That Old Janx Spirit
C	Tricia McMillan	Wine

File Edit View Run Kernel Tabs Settings Help											
Files	Code										
	Out[21]:	CampaignId	CampaignName	AdGroupId	AdGroupName	Date	Impressions	Clicks	Cost	Conversions	Converted
Running	0	10	Sentiment	100002	BANDAŽE & LĚKÁRNAČKA - Bandaje	2018-01-01	8	0	0.00	0	
	1	10	Sentiment	100002	BANDAŽE & LĚKÁRNAČKA - Bandaje	2018-01-02	10	0	0.00	0	
Commands	2	10	Sentiment	100002	BANDAŽE & LĚKÁRNAČKA - Bandaje	2018-01-03	23	2	7.80	0	
	3	10	Sentiment	100002	BANDAŽE & LĚKÁRNAČKA - Bandaje	2018-01-04	10	0	0.00	0	
	4	10	Sentiment	100002	BANDAŽE & LĚKÁRNAČKA - Bandaje	2018-01-05	8	0	0.00	0	
	5	10	Sentiment	100002	BANDAŽE & LĚKÁRNAČKA - Bandaje	2018-01-06	6	0	0.00	0	
Cell Tools	6	10	Sentiment	100002	BANDAŽE & LĚKÁRNAČKA - Bandaje	2018-01-07	4	0	0.00	0	
	7	10	Sentiment	100002	BANDAŽE & LĚKÁRNAČKA - Bandaje	2018-01-08	8	0	0.00	0	
Tabs	8	10	Sentiment	100002	BANDAŽE & LĚKÁRNAČKA - Bandaje	2018-01-09	5	0	0.00	0	
	9	10	Sentiment	100002	BANDAŽE & LĚKÁRNAČKA - Bandaje	2018-01-10	11	0	0.00	0	
	10	10	Sentiment	100002	BANDAŽE & LĚKÁRNAČKA - Bandaje	2018-01-11	5	0	0.00	0	
	11	10	Sentiment	100002	BANDAŽE & LĚKÁRNAČKA - Bandaje	2018-01-12	5	0	0.00	0	
	12	10	Sentiment	100002	BANDAŽE & LĚKÁRNAČKA - Bandaje	2018-01-13	5	0	0.00	0	
	13	10	Sentiment	100002	BANDAŽE & LĚKÁRNAČKA - Bandaje	2018-01-14	3	0	0.00	0	
	14	10	Sentiment	100002	BANDAŽE & LĚKÁRNAČKA - Bandaje	2018-01-15	15	1	2.85	0	
	15	10	Sentiment	100002	BANDAŽE & LĚKÁRNAČKA - Bandaje	2018-01-16	20	1	8.04	0	
	16	10	Sentiment	100002	BANDAŽE & LĚKÁRNAČKA - Bandaje	2018-01-17	30	4	15.30	0	
	17	10	Sentiment	100002	BANDAŽE & LĚKÁRNAČKA - Bandaje	2018-01-18	13	0	0.00	0	
	18	10	Sentiment	100002	BANDAŽE & LĚKÁRNAČKA - Bandaje	2018-01-19	13	1	4.45	0	
	19	10	Sentiment	100002	BANDAŽE & LĚKÁRNAČKA - Bandaje	2018-01-20	10	0	0.00	0	
	20	10	Sentiment	100002	BANDAŽE & LĚKÁRNAČKA - Bandaje	2018-01-21	6	0	0.00	0	
	21	10	Sentiment	100002	BANDAŽE & LĚKÁRNAČKA - Bandaje	2018-01-22	23	1	5.88	0	
	22	10	Sentiment	100002	BANDAŽE & LĚKÁRNAČKA - Bandaje	2018-01-23	35	2	7.58	0	

Return to Jupyter Notebook,
section Joining Tables



CONCLUSION

Other Features

- Working with Time Series and Time Deltas.
- Categorical data.
- Advanced statistics.
- Other I/O formats:
 - CSV,
 - HDF5,
 - JSON,
 - Google BigQuery,
 - and more...

Related Projects

Data Visualization

- [matplotlib](#)
- [seaborn](#)
- [Bokeh](#)

Computations

- [Statsmodels](#)
- [scipy](#)
- [Dask](#) and [Blaze](#)

N-dimensional structures

- [xarray](#)

Machine Learning

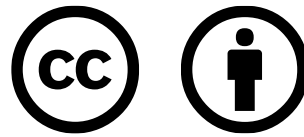
- [scikit-learn](#)
- [sklearn-pandas](#)

Output

- [pygsheets](#)



Any Questions?



Attribution 4.0 International (CC BY 4.0)

Thank you for listening!

We are looking for Python Devs!

<https://lynt.cz/kariera>

Interested in PPC?

Follow: [@PPCrobot](#)

