Data Analysis in Python: pandas in Practice

Jan Smitka

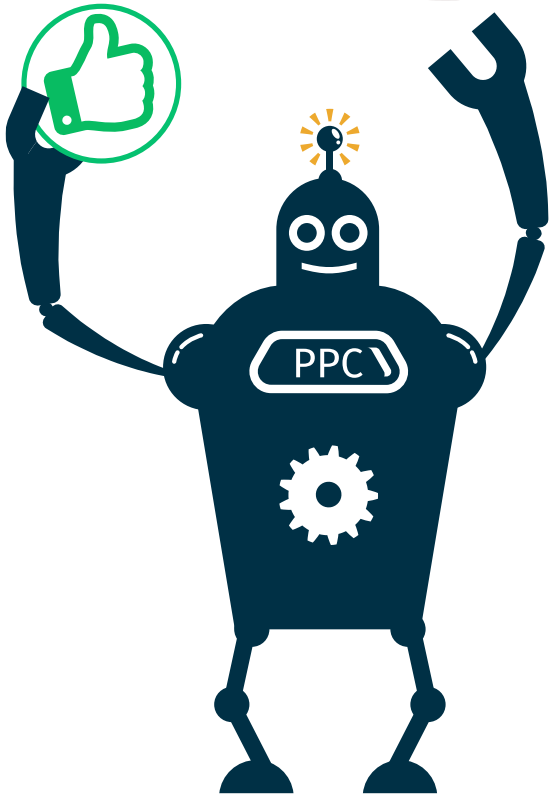
jan.smitka@lynt.cz

@jansmitka

Lynt services s.r.o.

 https://lynt.cz  @jansmitka Slides: https://u.lynt.cz/pandas

About Me




Marcador
Laurelin
Destructor
Azraell
arius Vicari
Pandas
heh
>-CuadrillerO-<
Crazynka
-=(Warcryer)=-
Creks
mala de San
Miriam
Sacerodisa
fishinka

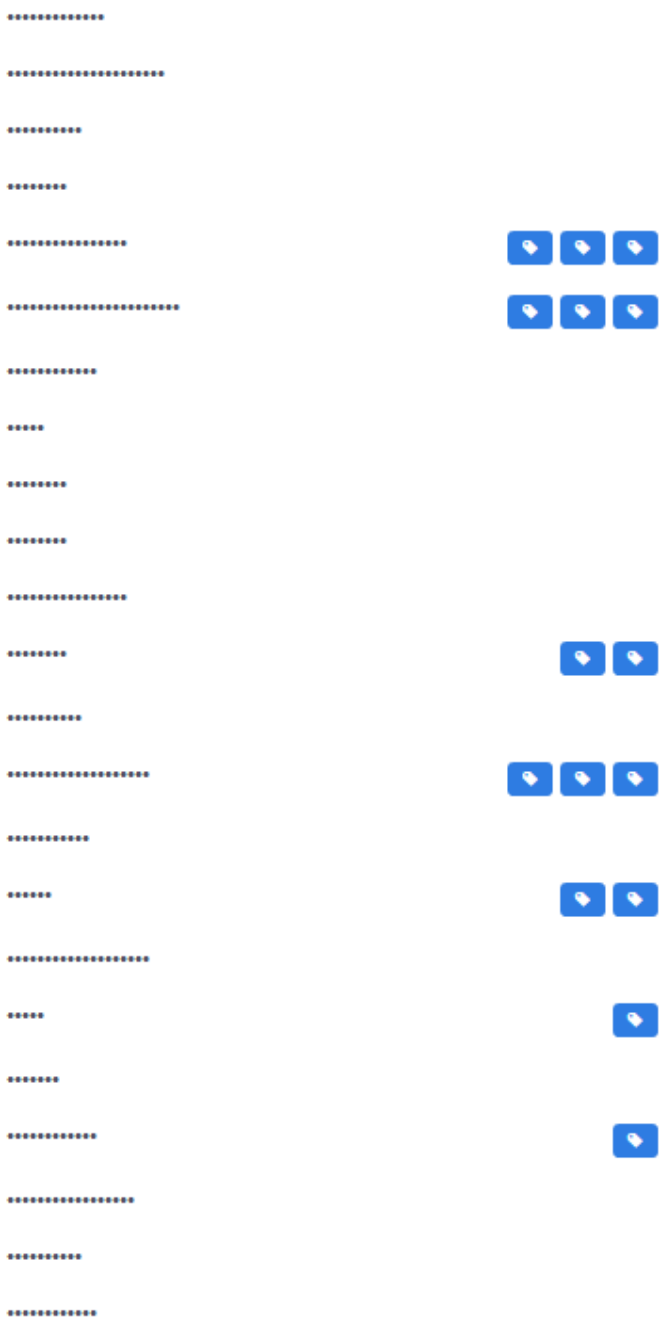
No need to write anything, everything is here:

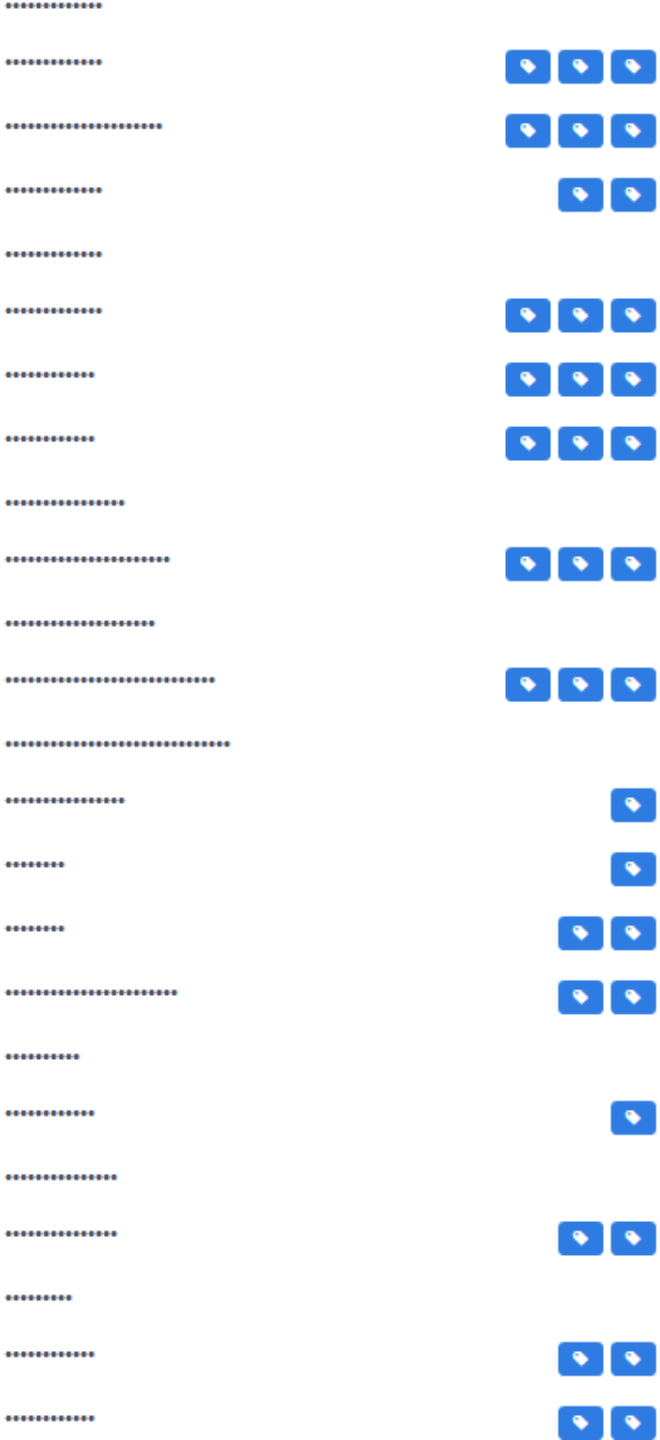
https://u.lynt.cz/pandas

MOTIVATION

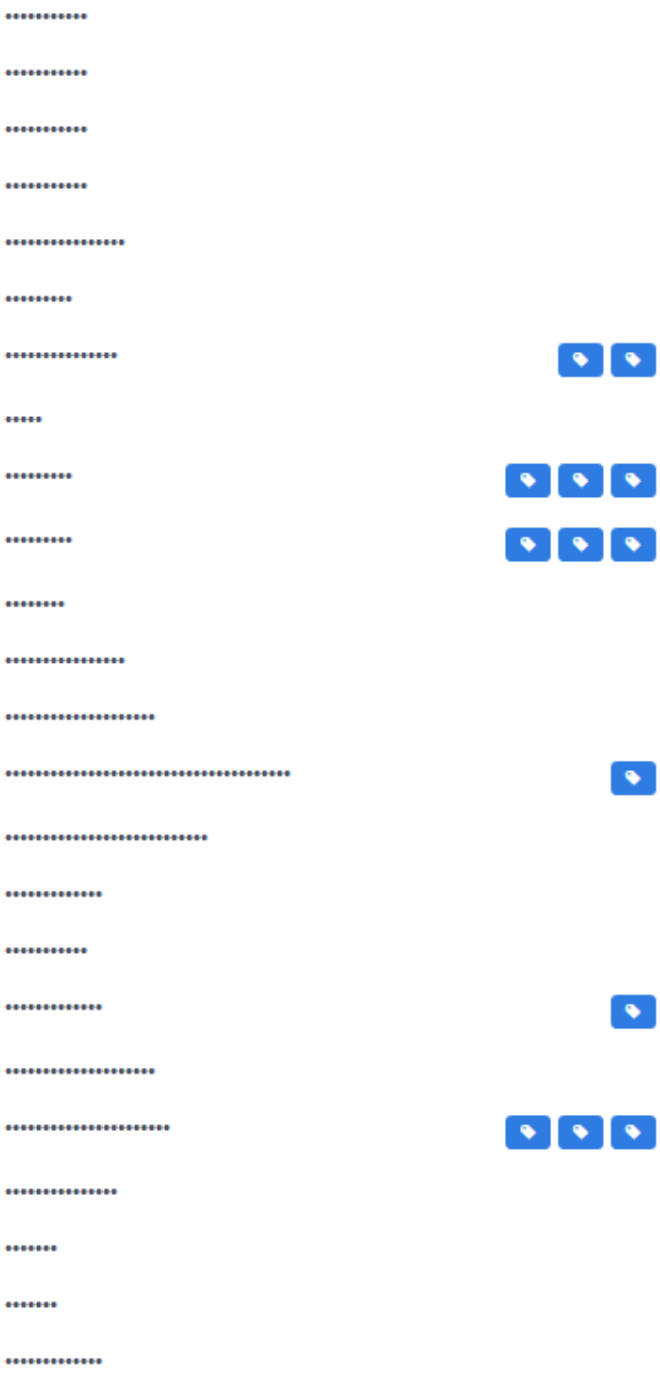
Why not just use Excel, Google Sheets, etc.?

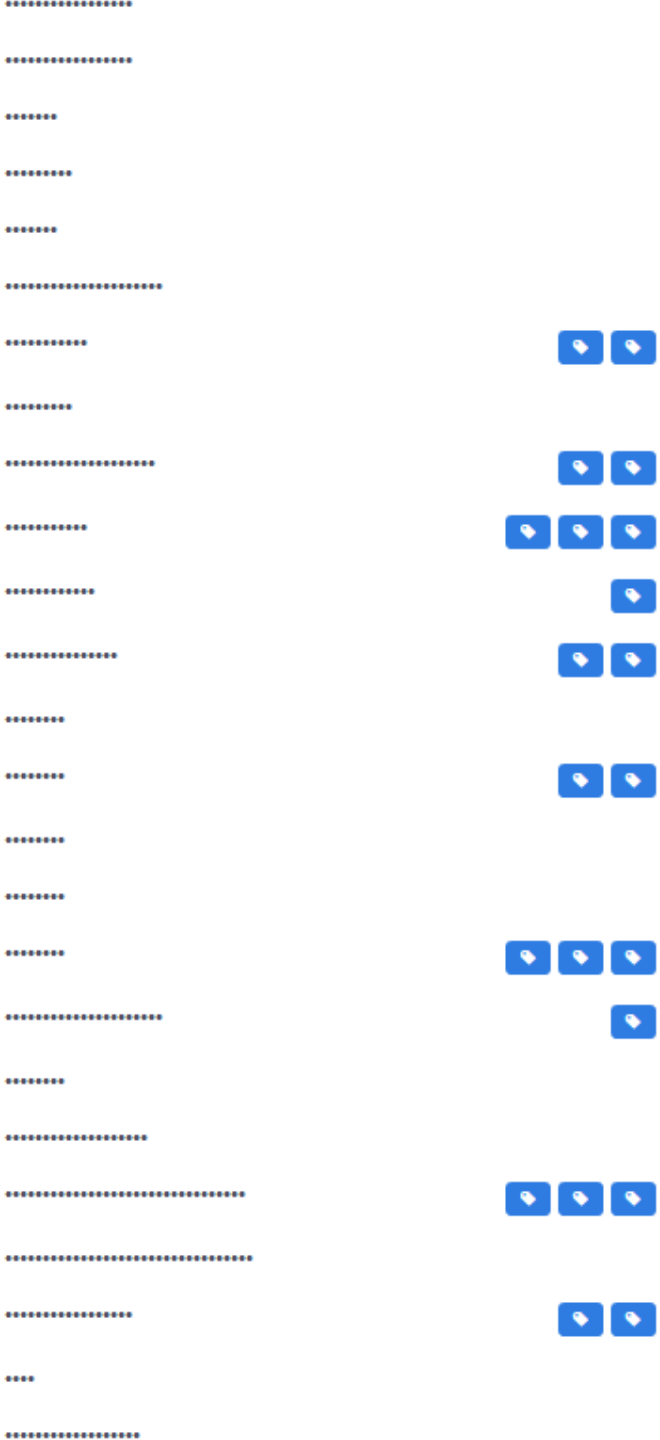
account-example@gmail.com



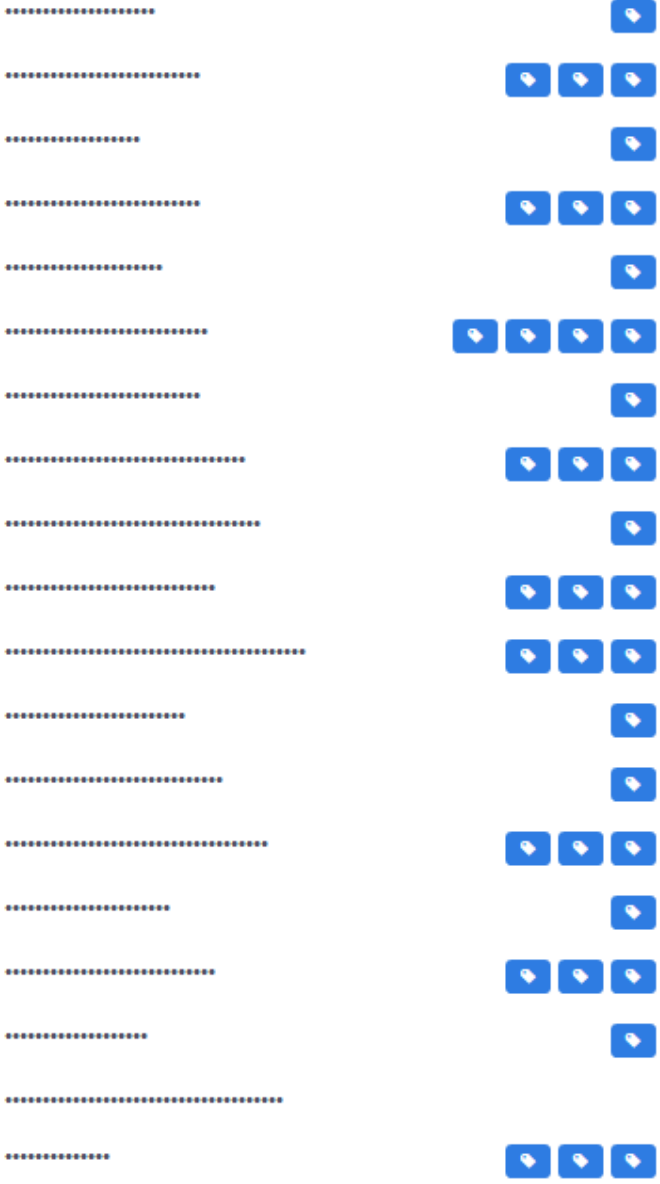


Quite a list!

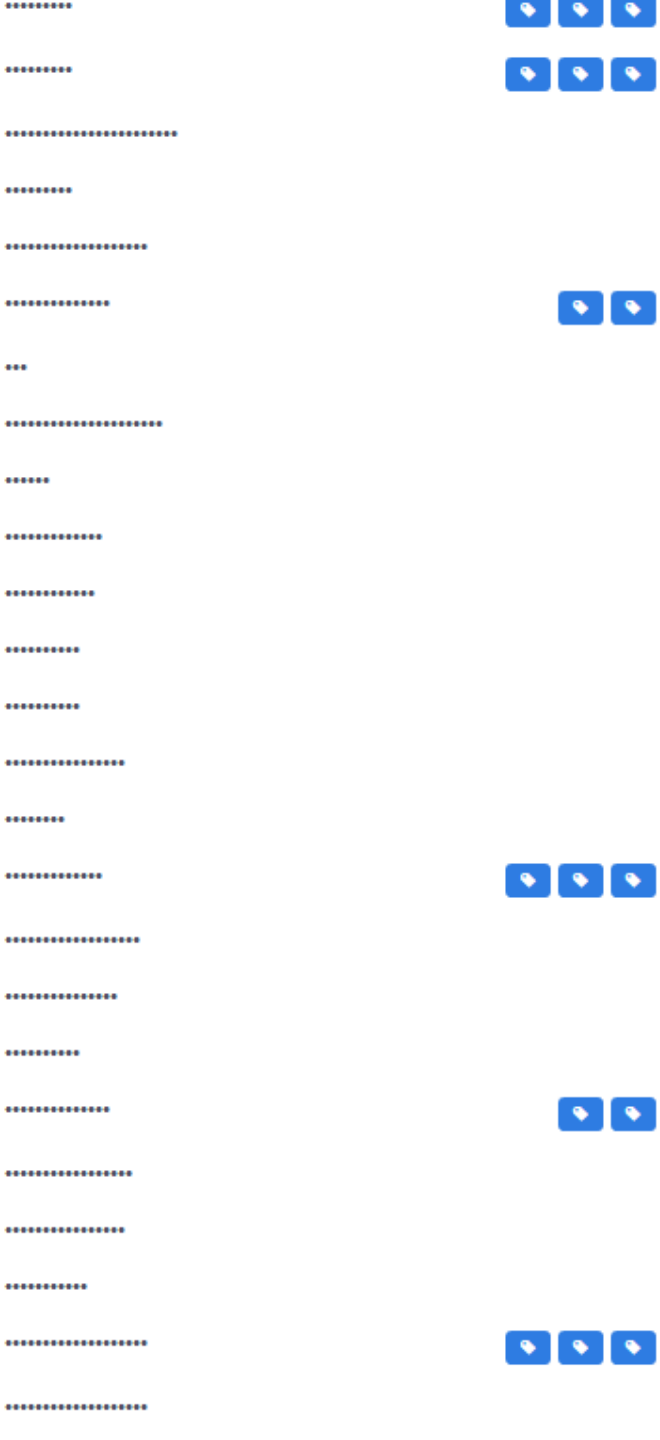


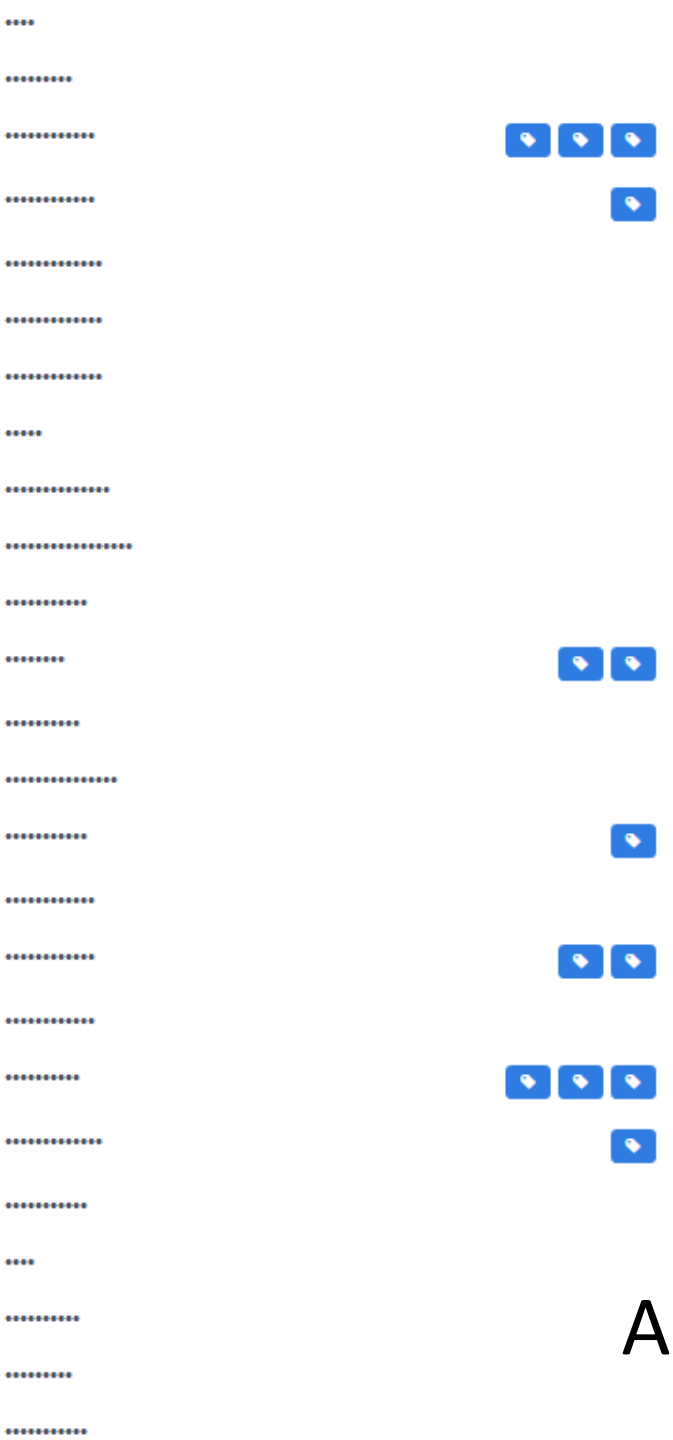


Still going…



Are we there yet?





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

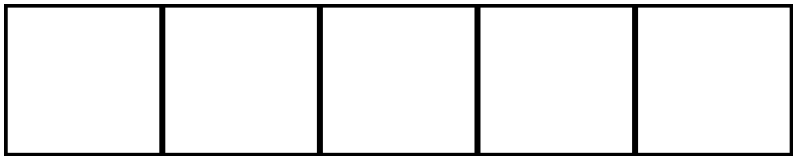
```
pip install pandas
```

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

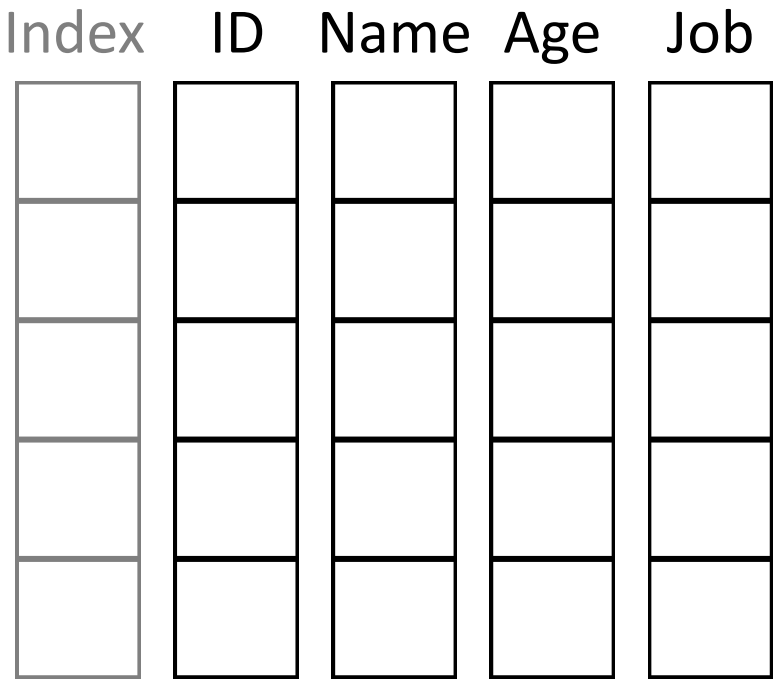
`pandas.Series`



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

Data structures: DataFrame

`pandas.DataFrame`



Collection of named Series with index.

Data types (dtype)

fixed-size numerics:

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

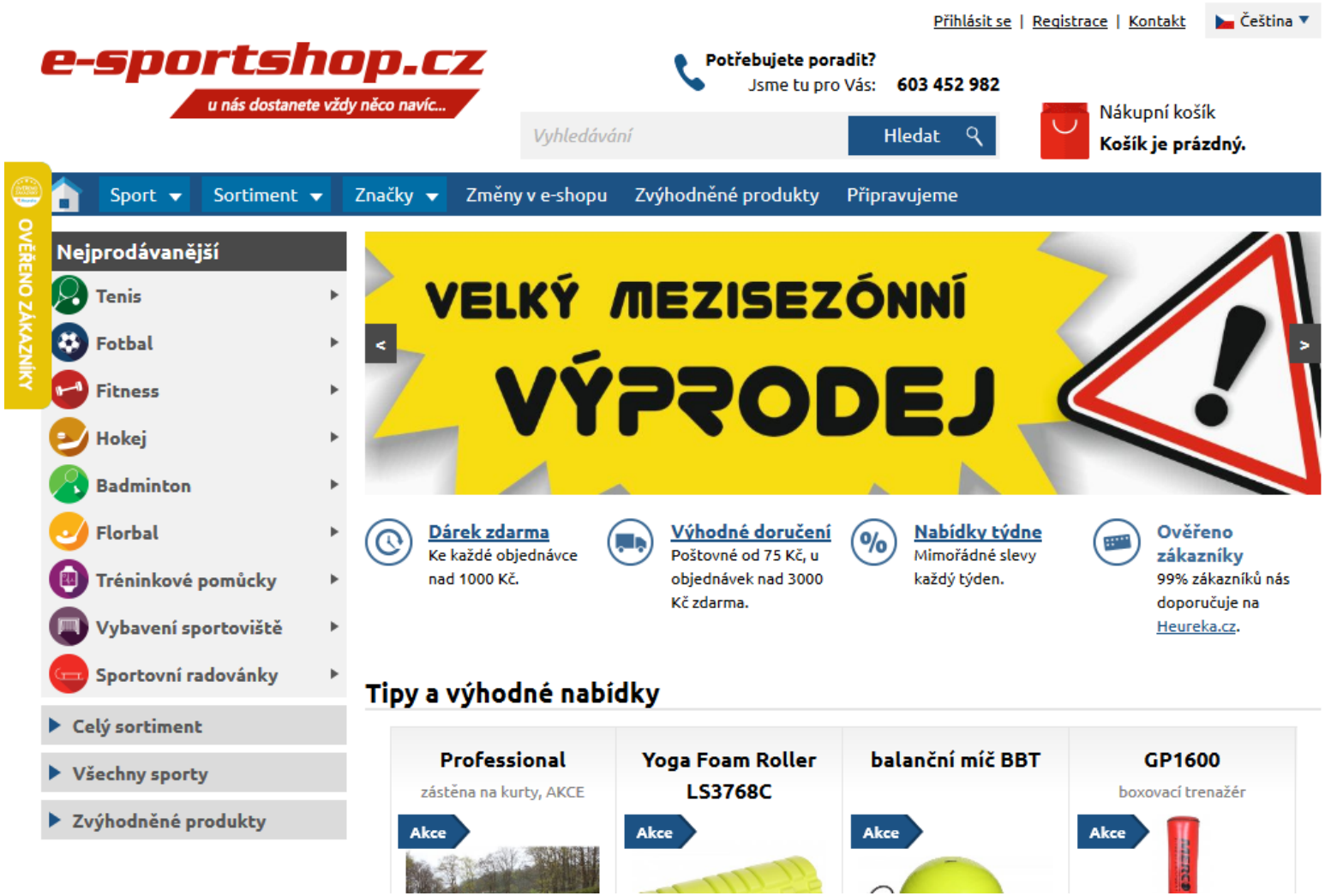
arbitrary data:

- category (like enum)
- object

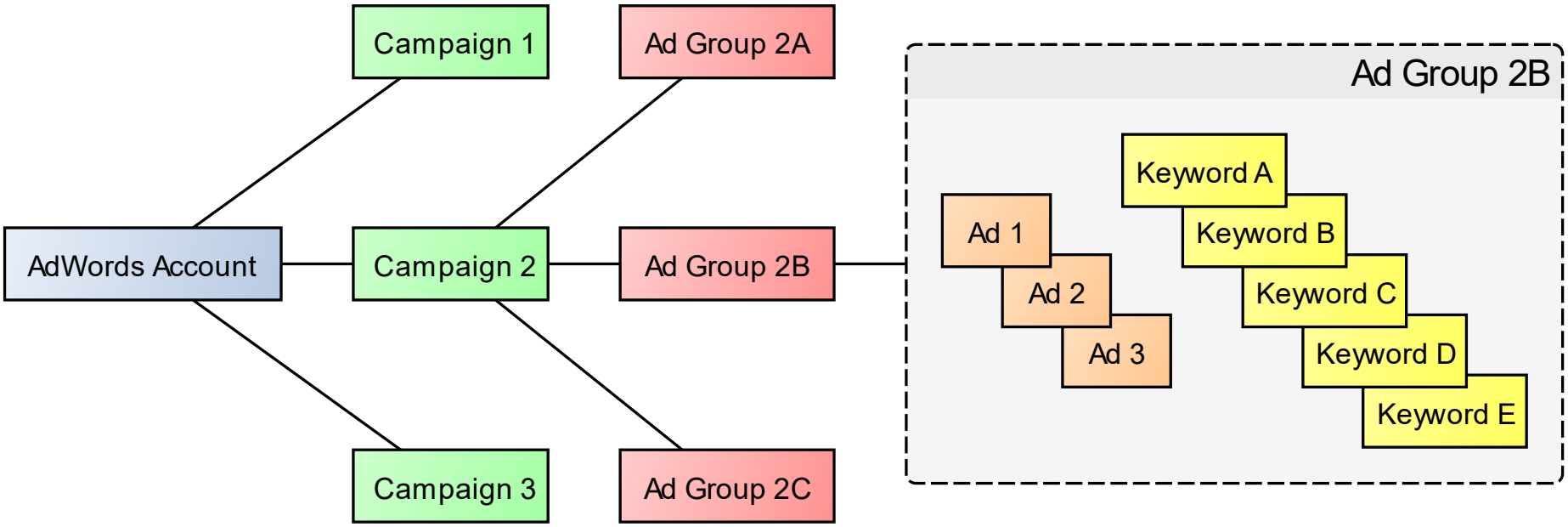
integers have limited size!

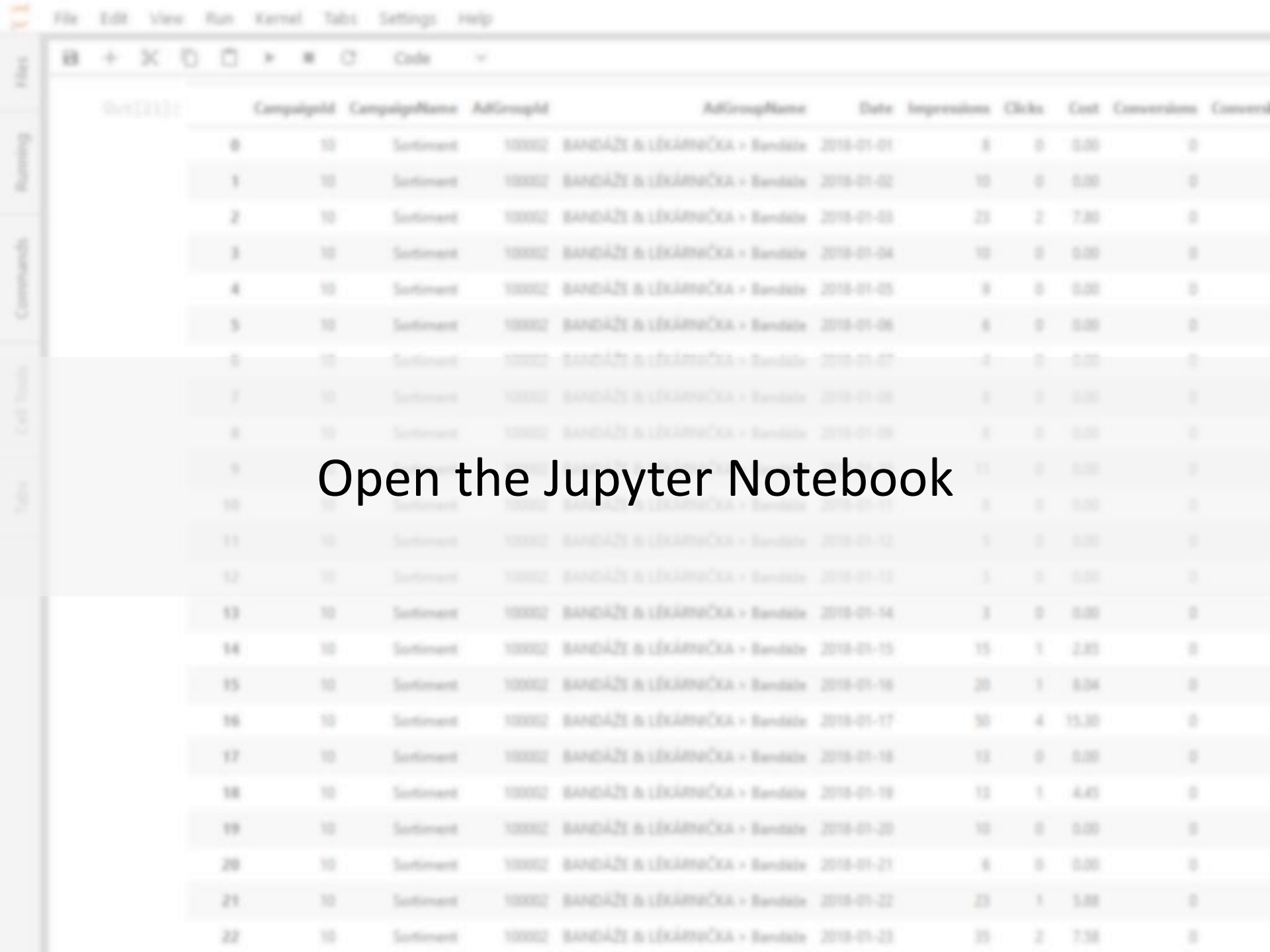
A REAL WORLD EXAMPLE

Source of Fake Data



AdWords Account Structure




Open the Jupyter Notebook

Common Metrics in AdWords

CTR (Click-Through Rate):

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

CPC (Cost Per Click):

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

Average Conversion Value:

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

Return to Jupyter Notebook, section Computations

Lynt SERVICES

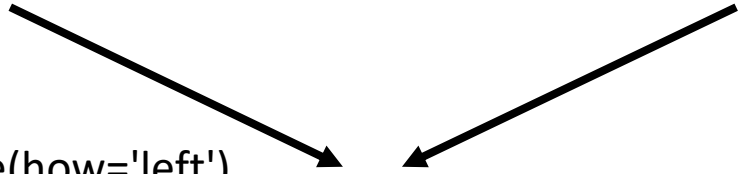
SQL-like joins: left join

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

Lynt SERVICES

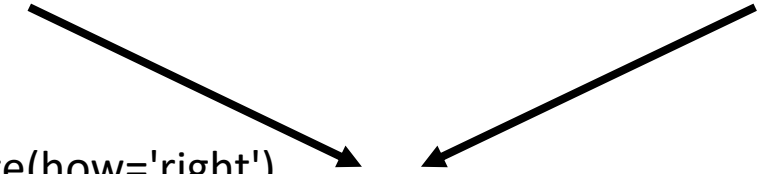
SQL-like joins: right join

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

Lynt SERVICES

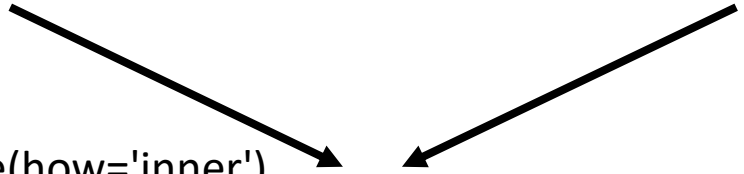
SQL-like joins: inner join

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

Lynt SERVICES

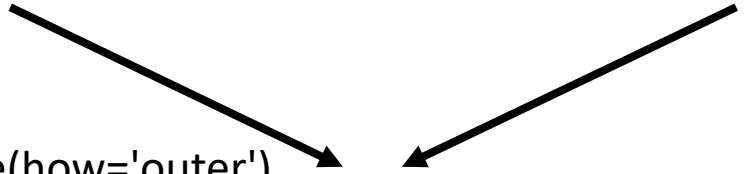
SQL-like joins: outer join

left

key	name
A	Arthur Dent
B	Ford Prefect
C	Tricia McMillan

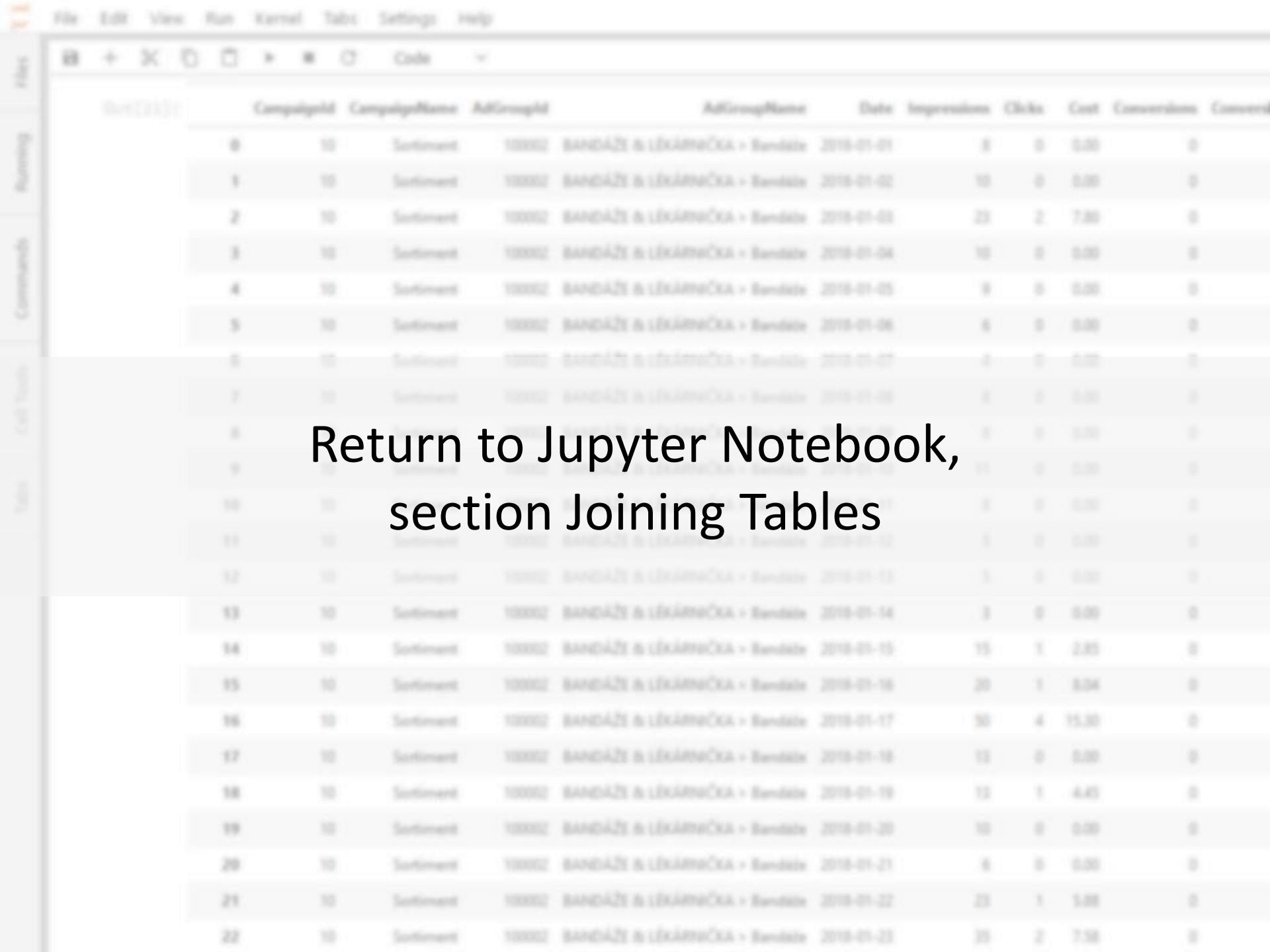
right

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


Return to Jupyter Notebook,
section Joining Tables

CONCLUSION

Other Features

- Working with Time Series and Time Deltas.
- Categorical data.
- Advanced statictics.
- 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

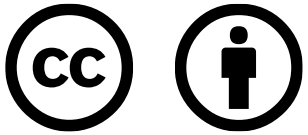
Machine Learning

- scikit-learn
- sklearn-pandas

Output

- pygsheets

Any Questions?



Attribution 4.0 International (CC BY 4.0)

Lynt SERVICES

Thank you for listening!

We are looking for Python Devs!

https://lynt.cz/kariera

Interested in PPC?

Follow: @PPCrobot