


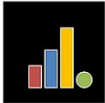

# Top 15 Python Tips for Data Cleaning/ Understanding

With two bonus tips!

By: Hui Xiang Chua

# A Little About Me



- Helps develop data architecture and analytics capabilities at **essence** 
- Runs a data science blog called  DATA DOUBLE CONFIRM
- Holds a B.Sc.(Hons) in Statistics/ M.Sc. in Business Analytics from 

# Tasks

1. Get column names
2. Get size of dataset
3. Check data type of variables
4. Get unique values
5. Get range of values
6. Get count of values
7. Rename column names
8. Remove symbols in values
9. Convert string to numeric/ string to date
10. Replace values with another value

# Tasks

11. Identify data variables (i.e. column names) similar/ different across datasets
12. Concatenate/ Appending
13. Deduplication
14. Merge
15. Recoding

[BONUS] 16. Data profiling

[BONUS] 17. Input missing values

# Use case

Data from various sources:

(1) campaign details, (2) viewability metrics, (3) brand lift study results, (...)

Common issues:

- inconsistent naming of variables/ fields across datasets
- inappropriate data formats
- invalid/ duplicate/ missing values

# Materials for today's talk



<https://tinyurl.com/y5b3y7to>

- Datasets
- Slides
- Jupyter notebook

# Necessary libraries

```
import pandas as pd
```

```
import numpy as np
```

What should I learn first pandas or NumPy? ^

**Numpy** provides the support of highly optimized multidimensional arrays, which are the most basic data structure of most Machine **Learning** algorithms. Next, you should **learn Pandas**. Data scientists spend most of their time cleaning data, which is also called as data munging or data wrangling.

[www.kdnuggets.com](#) › 2019/06 › python-data-science-rig...

[How to Learn Python for Data Science the Right Way - KDnuggets](#)

Is pandas based on NumPy? ^

**Pandas.** **pandas** is an open-source library built on top of **numpy** providing high-performance, easy-to-use data structures and data analysis tools for the Python programming language. It allows for fast analysis and data cleaning and preparation.

Nov 16, 2019

[towardsdatascience.com](#) › top-python-libraries-numpy-pa...

[Top Python Libraries: Numpy & Pandas - Towards Data Science](#)

Search for: [Is pandas based on NumPy?](#)

# Read in datasets

```
campaigns = pd.read_csv('mock_data_campaign.csv')  
metrics_h1 = pd.read_csv('mock_data_metrics_h1.csv')  
metrics_h2 = pd.read_csv('mock_data_metrics_h2.csv')
```



# Preview datasets

```
campaigns.head()
```

	campaign_id	team	vertical	market	channel	campaign_name	start_date	end_date	spends
0	2197	B	Music	IN	Twitter	B_Music_IN_Twitter	1/4/2019	10/13/2019	\$62,054
1	5577	C	Product	ID	Facebook	C_Product_ID_Facebook	1/16/2019	11/27/2019	\$59,945
2	3221	A	Music	JP	Facebook	A_Music_JP_Facebook	1/20/2019	4/17/2019	\$11,321
3	4339	D	Festive	ID	FB	D_Festive_ID_FB	1/28/2019	3/22/2019	\$79,436
4	7508	D	Festive	ID	OTT	D_Festive_ID_OTT	1/28/2019	3/21/2019	\$24,373

# Preview datasets

```
metrics_h1.head()
```

	campaign	impressions
0	B_Music_IN_Twitter	867976
1	C_Product_ID_Facebook	111888
2	A_Music_JP_Facebook	151285
3	D_Festive_ID_FB	752900
4	D_Festive_ID_OTT	580887

```
metrics_h2.head()
```

	campaign	impressions	measurable_impressions	clicks
0	B_Music_IN_YouTube	730769	720360	418046
1	C_Product_TH_OTT	162106	154224	46346
2	A_Music_KR_Facebook	11983	11980	9156
3	C_Product_ID_YouTube	52238	51938	13820
4	D_Festive_ID_DV	807004	799411	311062



# Questions?



[linkedin.com/in/hui-xiang-chua/](https://www.linkedin.com/in/hui-xiang-chua/)  
[linkedin.com/company/essence](https://www.linkedin.com/company/essence)



[facebook.com/essenceglobal](https://www.facebook.com/essenceglobal)



[@essence\\_global](https://www.instagram.com/essence_global)



[@hxchuaruns](https://twitter.com/hxchuaruns)  
[@essenceglobal](https://twitter.com/essenceglobal)



[projectosyo.wixsite.com/datadoubleconfirm](https://projectosyo.wixsite.com/datadoubleconfirm)