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 NUS
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
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

```python
campaigns.head()
```

<table>
<thead>
<tr>
<th>campaign_id</th>
<th>team</th>
<th>vertical</th>
<th>market</th>
<th>channel</th>
<th>campaign_name</th>
<th>start_date</th>
<th>end_date</th>
<th>spends</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>2197</td>
<td>B</td>
<td>IN</td>
<td>Twitter</td>
<td>B_Music_IN_Twitter</td>
<td>1/4/2019</td>
<td>10/13/2019</td>
<td>$62,054</td>
</tr>
<tr>
<td>1</td>
<td>5577</td>
<td>C</td>
<td>ID</td>
<td>Facebook</td>
<td>C_Product_ID_Facebook</td>
<td>1/16/2019</td>
<td>11/27/2019</td>
<td>$59,945</td>
</tr>
</tbody>
</table>
## Preview datasets

**metrics_h1.head()**

<table>
<thead>
<tr>
<th>campaign</th>
<th>impressions</th>
</tr>
</thead>
<tbody>
<tr>
<td>B_Music_IN_Twitter</td>
<td>867976</td>
</tr>
<tr>
<td>C_Product_ID_Facebook</td>
<td>111888</td>
</tr>
<tr>
<td>A_Music_JP_Facebook</td>
<td>151285</td>
</tr>
<tr>
<td>D_Festive_ID_FB</td>
<td>752900</td>
</tr>
<tr>
<td>D_Festive_ID_OTT</td>
<td>580887</td>
</tr>
</tbody>
</table>

**metrics_h2.head()**

<table>
<thead>
<tr>
<th>campaign</th>
<th>impressions</th>
<th>measurable_impressions</th>
<th>clicks</th>
</tr>
</thead>
<tbody>
<tr>
<td>B_Music_IN_YouTube</td>
<td>730769</td>
<td>720360</td>
<td>418046</td>
</tr>
<tr>
<td>C_Product_TH_OTT</td>
<td>162106</td>
<td>154224</td>
<td>46346</td>
</tr>
<tr>
<td>A_Music_KR_Facebook</td>
<td>11983</td>
<td>11980</td>
<td>9156</td>
</tr>
<tr>
<td>C_Product_ID_YouTube</td>
<td>52238</td>
<td>51938</td>
<td>13820</td>
</tr>
<tr>
<td>D_Festive_ID_DV</td>
<td>807004</td>
<td>799411</td>
<td>311062</td>
</tr>
</tbody>
</table>
Questions?

linkedin.com/in/hui-xiang-chua/
linkedin.com/company/essence

facebook.com/essenceglobal

@hxchuaruns
@essenceglobal

@essence_global

projectosyo.wixsite.com/datadoubleconfirm