# Python - Analiza danych z modułem PANDAS

www.udemy.com (http://www.udemy.com) (R)

# LAB - S04-L006-Przygotowanie danych

1. Zaimportuj moduł pandas i numpy, nadaj im standardowe aliasy. Do zmiennej **fb** wczytaj zawartość pliku **mrbean_facebook_statuses_with_nulls.csv**. Pobierz wszystkie kolumny. Wyświetl nagłówek obiektu data frame.
2. Sprawdź jakie kolumny ma **fb**, jakiego typu te kolumny są i ile pamięci zajmuje obiekt.
3. Ponownie wczytaj plik, ale teraz wybierz tylko kolumny: **"status_message", "status_type", "link_name", "num_reactions", "num_shares", "num_likes"**
4. Sprawdź jakie kolumny ma **fb**, jakiego typu te kolumny są i ile pamięci zajmuje obiekt.
5. Próbujemy optymalizować napisy:

    - Sprawdź ile wierszy ma **fb**
    - Sprawdź ile unikalnych wartości występuje w kolumnie **status_type**
    - Wyświetl te unikalne wartości wraz z informacją o ilości powtórzeń
    - Zmień typ na **category**
    - Sprawdź ile unikalnych wartości występuje w kolumnie **link_name**
    - Wyświetl te unikalne wartości wraz z informacją o ilości powtórzeń
    - Zmień typ na **category**

6. Sprawdź jakie kolumny ma **fb**, jakiego typu te kolumny są i ile pamięci zajmuje obiekt.
7. Próbujemy zoptymalizować liczby:

    - Zamień wartości NaN na 0 w kolumnach **num_reactions** i **num_shares**
    - Zmień typ w kolumnach **num_reactions, num_shares i num_likes** na int

8. Sprawdź jakie kolumny ma **fb**, jakiego typu te kolumny są i ile pamięci zajmuje obiekt.

# Rozwiązania:

Poniżej znajdują się propozycje rozwiązań zadań. Prawdopodobnie istnieje wiele dobrych rozwiązań, dlatego jeżeli rozwiązujesz zadania samodzielnie, to najprawdopodobniej zrobisz to inaczej, może nawet lepiej :) Możesz pochwalić się swoimi rozwiązaniami w sekcji Q&A

In [1]:
```
import pandas as pd
import numpy as np
fb = pd.read_csv("mrbean_facebook_statuses_with_nulls.csv")
fb.head()
```

Out[1]:

| | status_id | status_message | link_name | status_type | status_link | statu |
|---|---|---|---|---|---|---|
| **0** | 17774451468_10154154735571469 | It's time for Mr Bean and Teddy to get ready f... | Mr Bean - Preparing To Go Camping | video | https://youtu.be/fgURU75gTMQ | |
| **1** | 17774451468_10154146584106469 | NaN | Timeline Photos | photo | https://www.facebook.com/MrBean/photos/a.10150... | |
| **2** | 17774451468_10154135502911469 | NaN | Timeline Photos | photo | https://www.facebook.com/MrBean/photos/a.10150... | |
| **3** | 17774451468_10154138120151469 | Mr Bean is ready to do some shopping but bewar... | www.youtube.com | video | https://www.youtube.com/watch?v=58Z8J0PbLS8 | |
| **4** | 17774451468_10154135502476469 | NaN | Timeline Photos | photo | https://www.facebook.com/MrBean/photos/a.10150... | |

In [2]:
```
fb.info(memory_usage='deep')
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 56 entries, 0 to 55
Data columns (total 15 columns):
status_id           56 non-null object
status_message      40 non-null object
link_name           56 non-null object
status_type         56 non-null object
status_link         56 non-null object
status_published    56 non-null object
num_reactions       54 non-null float64
num_comments        55 non-null float64
num_shares          56 non-null float64
num_likes           56 non-null int64
num_loves           56 non-null int64
num_wows            56 non-null int64
num_hahas           56 non-null int64
num_sads            56 non-null int64
num_angrys          56 non-null int64
dtypes: float64(3), int64(6), object(6)
memory usage: 34.3 KB
```

In [3]:
```
fb = pd.read_csv("mrbean_facebook_statuses_with_nulls.csv",usecols=["status_message","
                                                                    "num_reactions","num_shares","
fb.head()
```

Out[3]:

| | status_message | link_name | status_type | num_reactions | num_shares | num_likes |
|---|---|---|---|---|---|---|
| **0** | It's time for Mr Bean and Teddy to get ready f... | Mr Bean - Preparing To Go Camping | video | 16570.0 | 338.0 | 16079 |
| **1** | NaN | Timeline Photos | photo | 119886.0 | 1657.0 | 114008 |
| **2** | NaN | Timeline Photos | photo | NaN | 10050.0 | 218579 |
| **3** | Mr Bean is ready to do some shopping but bewar... | www.youtube.com | video | 20913.0 | 257.0 | 20338 |
| **4** | NaN | Timeline Photos | photo | 223602.0 | 2228.0 | 215591 |

In [4]:
```
fb.info(memory_usage='deep')
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 56 entries, 0 to 55
Data columns (total 6 columns):
status_message    40 non-null object
link_name         56 non-null object
status_type       56 non-null object
num_reactions     54 non-null float64
num_shares        56 non-null float64
num_likes         56 non-null int64
dtypes: float64(2), int64(1), object(3)
memory usage: 16.0 KB
```

In [5]:
```
len(fb)
```

Out[5]: 56

In [6]:
```
fb["status_type"].nunique()
```

Out[6]: 3

In [7]:
```
fb["status_type"].value_counts()
```

Out[7]:
```
video    32
photo    22
link      2
Name: status_type, dtype: int64
```

In [8]:
```
fb["status_type"] = fb["status_type"].astype('category')
```

In [9]:
```
fb["link_name"].nunique()
```

Out[9]: 32

In [10]:
```
fb["link_name"].value_counts().head()
```

Out[10]:
```
Timeline Photos               22
www.youtube.com                4
Mr. Bean - Counting Sheep.     1
Mr Bean - Ray Of Sunshine      1
Mr Bean - Sun Block            1
Name: link_name, dtype: int64
```

In [11]:
```
fb["link_name"] = fb["link_name"].astype('category')
```

In [12]:
```
fb.info(memory_usage='deep')
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 56 entries, 0 to 55
Data columns (total 6 columns):
status_message    40 non-null object
link_name         56 non-null category
status_type       56 non-null category
num_reactions     54 non-null float64
num_shares        56 non-null float64
num_likes         56 non-null int64
dtypes: category(2), float64(2), int64(1), object(1)
memory usage: 12.6 KB
```

In [13]:
```
fb["num_reactions"].fillna(0,inplace = True)
fb["num_shares"].fillna(0,inplace = True)
```

In [14]:
```
fb["num_reactions"] = fb["num_reactions"].astype('int')
fb["num_shares"] = fb["num_shares"].astype('int')
fb["num_likes"] = fb["num_likes"].astype('int')
```

In [15]:
```
fb.info(memory_usage='deep')
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 56 entries, 0 to 55
Data columns (total 6 columns):
status_message    40 non-null object
link_name         56 non-null category
status_type       56 non-null category
num_reactions     56 non-null int32
num_shares        56 non-null int32
num_likes         56 non-null int32
dtypes: category(2), int32(3), object(1)
memory usage: 11.9 KB
```