

# Python - Analiza danych z modulem PANDAS

[www.udemy.com](http://www.udemy.com) (<http://www.udemy.com>) (R)

## LAB - S03-L013-LAB Duplikaty w data frame

1. Zaimportuj moduł pandas oraz numpy i nadaj im standardowe aliasy. Do zmiennej **surveys** wczytaj zawartość pliku **StackOverflowDeveloperSurvey2018.csv**. Pobierz tylko następujące kolumny: **'Country', 'Gender', 'OperatingSystem'**. Skorzystaj z parametru `low_memory=False`. Wyświetl nagłówek tak utworzonego Data Frame.
2. Sprawdź czy kolumna **Country** zawiera wartości unikalne
3. Wyznacz wszystkie kraje, z jakich pochodzą uczestnicy ankiety. (Wystarczy jak wyświetlisz pierwszych kilka)
4. Policzyć ile tych krajów jest (raz pomijając wartość **NaN** a raz uwzględniając ją)
5. Utwórz serię wartości True/False **duplicatesKeepFirst** zawierającą informacje o tym czy wystąpienie kraju w **surveys** jest pierwsze czy kolejne.
6. Wyświetl pierwszych kilka ankiet, które pochodzą z różnych krajów (skorzystaj z **duplicatedKeepFirst**)
7. Ile krajów zostanie zwróconych w poprzednim punkcie? Nim uruchomisz odpowiednie polecenie, spróbuj przewidzieć tą wartość.
8. Usuń powtarzające się wiersze ze względu na wartości w kolumnach **'Country', 'OperatingSystem'**
9. Wyświetl pozostałe w dataframe **surveys** wiersze dotyczące Twojego kraju

## 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
surveys = pd.read_csv("StackOverflowDeveloperSurvey2018.csv",
                      usecols=['Country', 'Gender', 'OperatingSystem'],
                      low_memory=False)

surveys.head()
```

Out[1]:

	Country	OperatingSystem	Gender
0	Kenya	Linux-based	Male
1	United Kingdom	Linux-based	Male
2	United States	NaN	NaN
3	United States	Windows	Male
4	South Africa	Windows	Male

```
In [2]: surveys.Country.is_unique
```

Out[2]: False

```
In [3]: #surveys.Country.unique()
surveys.Country.unique()[:20]
```

```
Out[3]: array(['Kenya', 'United Kingdom', 'United States', 'South Africa',
              'Nigeria', 'India', 'Spain', 'Croatia', 'Netherlands', 'Israel',
              'Sweden', 'Chile', 'Australia', 'Greece', 'Poland', 'Belgium',
              'Argentina', 'Germany', 'Russian Federation', 'Indonesia'],
              dtype=object)
```

```
In [4]: surveys.Country.nunique()
```

```
Out[4]: 183
```

```
In [5]: len(surveys.Country.unique())
```

```
Out[5]: 184
```

```
In [6]: duplicatesKeepFirst = surveys.duplicated(subset="Country")
```

```
In [7]: surveys[~duplicatesKeepFirst].head()
```

```
Out[7]:
```

	Country	OperatingSystem	Gender
0	Kenya	Linux-based	Male
1	United Kingdom	Linux-based	Male
2	United States	NaN	NaN
4	South Africa	Windows	Male
7	Nigeria	Windows	Female

```
In [8]: len(surveys[~duplicatesKeepFirst])
```

```
Out[8]: 184
```

```
In [9]: surveys.drop_duplicates(subset=['Country', 'OperatingSystem'], inplace=True)
```

```
In [10]: surveys.query("Country == 'Poland'")
```

```
Out[10]:
```

	Country	OperatingSystem	Gender
30	Poland	Linux-based	Male
34	Poland	Windows	Male
232	Poland	NaN	NaN
854	Poland	MacOS	Male
72587	Poland	BSD/Unix	NaN