Python - Analiza danych z modułem PANDAS

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

LAB - S02-L016 - Więcej funkcji Data Series

1. Zaimportuj moduł **pandas** i nadaj mu standardowy alias.
2. Do zmiennej programmers wczytaj serię danych z pliku **StackOverflowDeveloperSurvey2018.csv**. Wczytując dane użyj arguemtu **low_memory=False**, **usecols='ConvertedSalary** i **squeeze=True**.
3. Wylicz średnią, medianę i odchylenie standardowe. Co można powiedzieć o tych danych?
4. Wylicz maksymalną wartość **Salary**.
5. Do zmiennej **fortune500** wczytaj dane z pliku **Fortune_500_2017.csv**. Wczytaj kolumny **Title** i **Employees**. Kolumna Title powinna stać się indeksem. Wyświetl nagłówek otrzymanej serii danych.
6. Wyświetl jaki jest indeks dla pozycji z największą liczbą pracowników. W ten sposób znajdziesz na liście **Fortune** firmę zatrudniającą najwięcej pracowników.
7. A ile pracowników zatrudnia ta firma. Skorzystaj z wartości wyliczonej w poprzednim punkcie
8. Wyświetl jaki jest indeks dla pozycji z najmniejszą liczbą pracowników. W ten sposób znajdziesz na liście **Fortune** firmę zatrudniającą najmniej pracowników
9. A ile pracowników zatrudnia ta firma. Skorzystaj z wartości wyliczonej w poprzednim punkcie

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
```

In [2]:
```
programmers = pd.read_csv("StackOverflowDeveloperSurvey2018.csv",
                          low_memory=False,usecols=['ConvertedSalary'],squeeze=True).c
```

In [3]:
```
programmers.mean()
```

Out[3]:
```
95780.86178776571
```

In [4]:
```
programmers.median()
```

Out[4]:
```
55075.0
```

In [5]:
```
programmers.std()
```

Out[5]:
```
202348.21562528735
```

Dane są "dziwne". O ile można uwierzyć w średnią i medianę, o tyle odchylenie standardowe wskazuje, że mamy dużo skrajnych wartości, co może świadczyć o nierzetelności informacji...

In [6]:
```
programmers.max()
```

Out[6]:
```
2000000.0
```

In [7]:
```
fortune = pd.read_csv("Fortune_500_2017.csv", usecols=['Employees','Title'],index_col=
fortune.head()
```

Out[7]:

Title	Employees
Walmart	2300000
Berkshire Hathaway	367700
Apple	116000
Exxon Mobil	72700
McKesson	68000

In [8]:
```
fortune.idxmax()
```

Out[8]:
```
Employees    Walmart
dtype: object
```

In [9]:
```
fortune.loc[fortune.idxmax()]
```

Out[9]:

Title	Employees
Walmart	2300000

In [10]:
```
fortune.idxmin()
```

Out[10]:
```
Employees    A-Mark Precious Metals
dtype: object
```

In [11]:
```
fortune.loc[fortune.idxmin()]
```

Out[11]:

Title	Employees
A-Mark Precious Metals	83