Очистка и подготовка данных

Data pre-processing and visualization

Азамат Хасенов

Очистка и подготовка данных

- В реальном мире данные обычно «**грязные**».
- Например:

G15

	A	B	C	D	E	F
1	id	gender	nationality	city	data of birth	
2	1	(male)	[Kazakhstan]	Aktobe	1990	
3	2	Female	russia	moscow	1956-1957	
4	3	male	kazakhstan	astana	2001	
5	4	FEMALE	Kazakhstan	nur-sultan	2019	
6	5	male	Russia	Saint Petersb	1999	
7	6	female)	ukraine	Kiev	1989	
8	7		kazakhstan	almaty	2008	
9	8	Male	korea	pusan	1999	
10	9	(female)	korea	busan	1980	
11						

Очистка и подготовка данных

```
import csv
```

```
with open("synthetic_data.csv","r") as file:
    csv_data=csv.reader(file)
    data=list(csv_data)
    for row in data:
        print(row)
```

```
['id', 'gender', 'nationality', 'city', 'data of birth']
['1', '(male)', '[Kazakhstan]', 'Aktobe', '1990']
['2', 'Female', 'russia', 'moscow', '1956-1957']
['3', 'male', 'kazakhstan', 'astana', '2001']
['4', 'FEMALE', 'Kazakhstan', 'nur-sultan', '2019']
['5', '   male', 'Russia', 'Saint Petersburg', '1999']
['6', 'female)', 'ukraine', 'Kiev', '1989']
['7', '', 'kazakhstan', 'almaty', '2008']
['8', 'Male', 'korea', 'pusan', '1999']
['9', '(female)', 'korea', 'busan', '1980']
```

Очистка и подготовка данных

```
gender = "(male)"
print(f"Before: {gender}")
gender = gender.replace("(","")
print(f"First change: {gender}")
gender = gender.replace(")","")
print(f"Second change: {gender}")
```

```
Before: (male)
First change: male)
Second change: male
```

Очистка и подготовка данных

```
nationality="kazakhstan"
print(f"Before: {nationality}")
nationality=nationality.title()
print(f"Before: {nationality}")
```

```
Before: kazakhstan
Before: Kazakhstan
```

Очистка и подготовка данных

```
birth_year = "1956-1958"

def average_year(year):
    if "-" in year:
        years_str = year.split("-")
        years_int = [int(y) for y in years_str]
        app_year = round((years_int[0]+years_int[1])/2)
        year=app_year
    return int(year)
```

```
average_year(birth_year)
```

```
1957
```

Азамат Хасенов

Очистка и подготовка данных

```
exString = "   Hello!  "

def remove_front_spaces(inputString):
    return inputString.lstrip()

def remove_back_spaces(inputString):
    return inputString.rstrip()

def remove_both_spaces(inputString):
    return inputString.strip()

f1 = remove_front_spaces(exString)
f2 = remove_back_spaces(exString)
f3 = remove_both_spaces(exString)
```

Очистка и подготовка данных

```
f1
```

```
'Hello!   '
```

```
f2
```

```
'   Hello!'
```

```
f3
```

```
'Hello!'
```

Азамат Хасенов