
Pandas

Краткий план

1. Познакомиться с понятием DataFrame объекта в Pandas
2. Попробовать работу с Pandas в консоли
 - а. Откройте **jupyter-тетрадь** (notebook) и работайте в ней параллельно со слайдами
3. Практический урок в **Jupyter-тетради**

Что такое Pandas?

- Библиотека **структур данных** и инструментов для **анализа данных**:
 - Excel на Python
- Базовые объекты для работы с данными – это numpy-массивы
- Обратите внимание: если вы пользуетесь **популярной** библиотекой с большой базой пользователей (как например numpy), и у вас есть вопрос – **ответ на него скорее всего уже есть** на **StackOverflow**!
- **Документация** по Pandas отлично написана (на английском языке)
 - <https://pandas.pydata.org/pandas-docs/stable/10min.html>

«Анатомия» объекта DataFrame в Pandas

Diagram illustrating the structure of a DataFrame. The columns are labeled: InvoiceNo, CustomerID, Quantity, and UnitPrice. The rows are labeled: A, B, C, D, E, F, G, H, and I. A specific row (G) and column (CustomerID) are highlighted with a black border, and the intersection cell is labeled 'row' and 'column'.

| | InvoiceNo | CustomerID | Quantity | UnitPrice |
|---|-----------|------------|----------|-----------|
| A | | | | |
| B | | | | |
| C | | | | |
| D | | | | |
| E | | | | |
| F | | | | |
| G | | | | |
| H | | | | |
| I | | | | |

Diagram illustrating the structure of a DataFrame. The columns are labeled: InvoiceNo, CustomerID, Quantity, and UnitPrice. The rows are labeled: A, B, C, D, E, F, G, H, and I. A specific row (G) and column (CustomerID) are highlighted with a black border. The diagram also shows the underlying data types for each column: InvoiceNo is a `pd.Series`, CustomerID is a `np.array`, and Quantity and UnitPrice are `pd.DataFrame`.

| | InvoiceNo | CustomerID | Quantity | UnitPrice |
|---|-----------|------------|----------|-----------|
| A | | | | |
| B | | | | |
| C | | | | |
| D | | | | |
| E | | | | |
| F | | | | |
| G | | | | |
| H | | | | |
| I | | | | |

`pd.Series` `np.array` `pd.DataFrame`

Практическое введение

Откройте jupyter-тетрадь...

```
$ import pandas as pd
```

```
$ data = {'name': ['alina', 'oleg'], 'age': [23, 32]}
```

```
$ df = pd.DataFrame(data)
```

```
$ df
```

```
>  name  age  
0  alina  23  
1  oleg   32
```

Что такое DataFrame?

\$ df.values

```
> array([[ 'alina', 23],  
        [ 'oleg', 32]], dtype=object)
```

\$ df.dtypes

\$ df.columns

\$ df.index

\$ df.index = ['first', 'second']

\$ df

```
>      name  age  
first  alina  23  
second oleg   32
```

Индексация и выборка

```
$ df['name'] # или df.name
```

```
$ df.loc['second', 'age']
```

```
$ df.iloc[1, 1]
```

```
$ df.query('age < 30') > name age first alina 23
```

Чтение из CSV-файла

```
$ loc = 'data/your_data.csv'  
$ pd.read_csv(loc)
```

Можно читать с web-сайтов!

Другие аргументы, например `header` (шапка), позволяют настроить импорт: обращаться к колонкам по имени и выбрать колонку, по которой будет проходить индексация.

```
$ url = 'https://goo.gl/XE5CrW'  
$ pd.read_csv(url, header=None)
```

Прочтите [документацию](#)!

Полезные функции, методы и атрибуты

```
$ pd.isnull(df)
$ df.fillna(value=0)
$ df.describe()
$ df.plot()
$ df.reset_index()
$ df.set_index('name')
$ df.index
$ df.values # Симба, не забывай, кто ты...
$ df.col.unique() # а также математические функции вроде df.col.max()
$ df.groupby(...)
```



Практическое занятие

01-pandas-skeleton-rus.ipynb

20 минут