Тема 5. Pandas для анализа данных

Синтаксис

Вызов библиотеки pandas

```
import pandas
import pandas as pd
```

Конструктор DataFrame() для создания таблицы

```
pd.DataFrame(data = data, columns = columns)
# аргумент data - список с данными,
# аргумент columns - список с названиями столбцов
```

Метод `read_csv()` для чтения файлов формата CSV

```
df = pd.read_csv('путь к файлу')
```

Метод head() для вывода первых строк таблицы

```
df.head() # первые 5 строк
df.head(10) # первые 10 строк
```

Метод tail() для вывода последних строк таблицы

```
df.tail() # последние 5 строк
df.tail(15) # последние 15 строк
```

Атрибут columns для вывода названий столбцов

```
df.columns
```

Атрибут shape для вывода размера таблицы

```
df.shape
```

Атрибут dtypes для получения информации о типах данных в таблице

```
df.dtypes
```

Метод info() для просмотра сводной информации о таблице

```
df.info()
```

Атрибут loc[строка, столбец] даёт доступ к элементу в DataFrame по строке и столбцу.

```
df.loc[:, 'column']
```

Вид	Реализация
Одна ячейка	`.loc[7, 'column']`
Один столбец	`.loc[:, 'column']`
Несколько столбцов	`.loc[:, ['column_1', 'column_4']]`
Несколько столбцов подряд (срез)	`.loc[:, 'column_5': 'column_8']`
Одна строка	`.loc[1]`
Все строки, начиная с заданной	`.loc[1:]`
Все строки до заданной	`.loc[:3]`
Несколько строк подряд (срез)	`.loc[2:5]`

Логическая индексация для получения элементов по определенному условию.

Вид	Реализация	Сокращенная запись
Все строки, удовлетворяющие условию	‘df.loc[df.loc[:,’column’] == ’X’]‘	‘df[df[’column’] == ’X’]‘
Столбец, удовлетворяющий условию	‘df.loc[df.loc[:,’column’] == ’X’][’column’]‘	‘df[df[’column’] == ’X’][’column’]‘
Применение метода	‘df.loc[df.loc[:,’column’] == ’X’][’column’].count()‘	‘df[df[’column’] == ’X’][’column’].count()‘

Индексация в Series

Словарь

Библиотека – это набор готовых методов для решения распространенных задач.

CSV (от англ. Comma-Separated Values, «значения, разделённые запятой») — формат файла. Каждая строка представляет собой одну строку таблицы, где данные разделены запятыми. В первой строке собраны заголовки столбцов (если они есть).

Кортеж – одномерная неизменяемая последовательность данных. Она похожа на список, её тоже можно сохранять в переменной.

DataFrame — это двумерная структура данных **Pandas**, где у каждого элемента есть два индекса: по строке и по столбцу.

- Каждая строка — это одно наблюдение, запись об объекте исследования. А столбцы — признаки этого объекта.
- `DataFrame()` – это конструктор библиотеки Pandas, который используется для создания **DataFrame**. Перед именем конструктора стоит обращение к переменной, в которой библиотека хранится: `pd.DataFrame()`.
- У DataFrame есть неотъемлемые свойства, значения которых можно запросить. Они называются атрибуты. Например, это размер таблицы `df.shape` или количество столбцов `df.columns`.
- К каждой ячейке с данными в DataFrame можно обратиться по её индексу и названию столбца. Этот процесс называется **индексация** и для DataFrame его проводят разными способами.

Series — одномерная структура данных Pandas, её элементы можно получить по индексу. Каждый индекс представляет собой номер отдельного наблюдения, и поэтому несколько различных Series вместе составляют DataFrame.

- В Series хранятся данные одного типа.
- У Series есть имя (Name), информация о количестве данных в столбце (Length) и тип данных, которые хранятся в ней (dtype).
- Индексация в Series аналогична индексации элементов столбца в DataFrame.