БДЗ

«Анализ данных на Python»

Модуль pandas

pandas – это модуль, который предоставляет очень удобные с точки зрения использования инструменты для хранения данных и работе с ними. Если вы занимаетесь анализом данных или машинным обучением и при этом используете язык Python, то вы просто обязаны знать и уметь работать с *pandas*.

pandas позволяет строить сводные таблицы, выделять колонки, использовать фильтры, выполнять группировку по параметрам, запускать функции (сложение, нахождение медианы, среднего, минимального, максимального значений), объединять таблицы и многое другое. В *pandas* можно создавать и многомерные таблицы.

Документация: https://pandas.pydata.org/docs/

pandas не входит в стандартный набор модулей Python, для установки данного модуля необходимо в консоли выполнить следующую команду:

```
pip install pandas
```

pandas работает со следующими структурами данных:

1. Series
2. DataFrame
3. Panel

Series

Pandas Series – это проиндексированный одномерный массив значений. Он похож на простой словарь типа ***dict***, где имя элемента будет соответствовать индексу, а значение – значению записи.

> *Примечание.* Размер структуры данных серии в Pandas является неизменным, т. е. однажды установленный, он не может быть изменен динамически. При этом значения и элементы в серии можно изменять.

Пример

```
import pandas as pd

data = ['Red', 'Green', 'Blue']
index = [100, 101, 102]
series_data = pd.Series(data, index=index)
print(series_data)
```

Вывод

```
100      Red
101    Green
102     Blue
dtype: object
```

Ключевые моменты

1. Однородные данные
2. Размер неизменный
3. Значения изменяемых данных

DataFrame

pandas предоставляет DataFrame, который представляет собой двумерную структуру, напоминающую двумерные массивы. Здесь входные данные оформляются в виде строк и столбцов.

> *Примечание.* Размер структуры данных DataFrame в Pandas можно изменять.

Пример

```
import pandas as pd

data = [
    ['Apple', 'Red'],
    ['Pear', 'Green'],
    ['Orange', 'Orange']
]

data_frame = pd.DataFrame(data, columns=['Fruit', 'Color'], index=[1, 2, 3])
print(data_frame)
```

Вывод

```
        Fruit   Color
    1   Apple     Red
    2    Pear   Green
    3  Orange  Orange
```

Ключевые моменты

1. Гетерогенные данные
2. Размер изменчивый
3. Изменяемые данные

Panel

pandas предлагает панель, которая представляет собой трехмерную структуру данных и содержит 3 оси для выполнения следующих функций:

- items: (ось 0). Каждый его элемент соответствует DataFrame в нем.
- major_axis: (ось 1). Соответствует строкам каждого DataFrame.
- minor_axis: (ось 2). Соответствует столбцам каждого DataFrame.

Ключевые моменты

1. Гетерогенные данные
2. Размер изменчивый
3. Изменяемые данные

Структура данных	**Размер**	**Описание**
Series	1	1D однородный массив, размер не изменяемый.
DataFrame	2	Двумерная таблично-изменяемая структура с потенциально разнородными столбцами.
Panel	3	3D изменяемый по размеру массив.

Импорт данных из файла CSV в DataFrame

Модуль DataFrame Python Pandas также может быть построен с использованием файлов CSV. Файл CSV – это в основном текстовый файл, в котором хранятся данные для каждой строки. Элементы разделяются запятой. Метод read_csv (file_name) используется для чтения данных из файла CSV в DataFrame.

Пример

```
import pandas as pd
data = pd.read_csv('sample.csv')
print(data)
```

Вывод

```
     Fruit   Color
1   Apple     Red
2    Pear   Green
3  Orange  Orange
```

Статистический анализ

Модуль ***pandas*** предлагает большое количество встроенных методов, помогающих пользователям проводить статистический анализ данных:

https://pandas.pydata.org/pandas-docs/stable/reference/frame.html#computations-descriptive-stats

Операции с текстовыми данными

Строковые функции Python можно применять к DataFrame.

Список используемых строковых функций в DataFrame:

https://pandas.pydata.org/pandas-docs/stable/reference/series.html#string-handling

Задания

Рассмотрим пример популярной задачи предсказания выживших пассажиров Титаника. Ознакомьтесь с файлом *train.csv*

Столбец	**Описание**	**Значения**
PassengerId	ID пассажира	
Survived	Выжил ли пассажир	0 = Нет, 1 = Да
Pclass	Класс билета	1 = 1st, 2 = 2nd, 3 = 3rd
Name	Имя пассажира	
Sex	Пол	male / female
Age	Возраст (в годах)	
SibSp	Братья / сестры на борту	
Parch	Родители / дети на борту	
Ticket	Номер билета	
Fare	Тариф	
Cabin	Номер каюты	
Embarked	Порт посадки	C = Cherbourg, Q = Queenstown, S = Southampton

Файл *test.csv* содержит аналогичные данные, кроме столбца *Survived*.

Задание №1

С помощью модуля ***pandas*** выведите статистику погибших/выживших отдельно для мужчин и женщин в каждом классе (*Pclass*).

Задание №2

С помощью модуля ***pandas*** выведите статистику по всем числовым полям, отдельно для мужчин и женщин.

Задание №3

Определите, влияет ли порт посадки на выживаемость.

Задание №4

4.1. Выведите топ 10 популярных имён.

4.2. Выведите топ 10 популярных фамилий.

Задание №5

Заполните все отсутствующие в *train.csv* значения медианой (*по столбцу*).

Задание №6

На основе статистики попытайтесь предсказать выживаемость для пассажиров из файла *test.csv*.

Задание №7 (3 балла)

Зарегистрируйтесь/авторизуйтесь на сайте https://www.kaggle.com

Загрузите своё решение на https://www.kaggle.com/c/titanic/

Задание №8 (+3 дополнительных балла)

С помощью библиотеки matplotlib отобразите гистограмму зависимости возраста (в годах) от выживаемости.