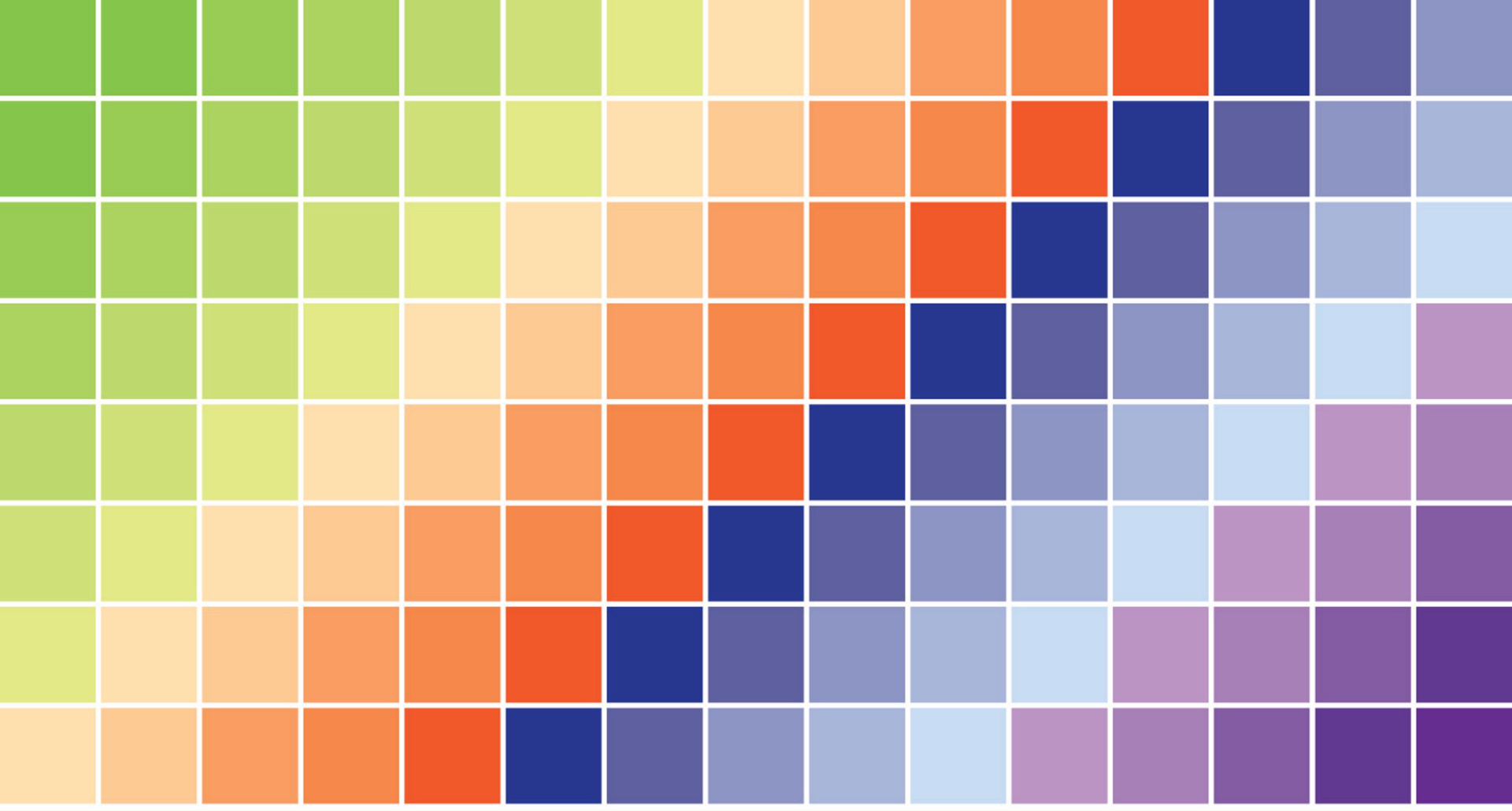


School of Computer Science

Библиотека анализа данных Pandas

ПРОГРАММИРОВАНИЕ НА PYTHON

Лекции для IT-школы

ЧТО ТАКОЕ АНАЛИЗ ДАННЫХ

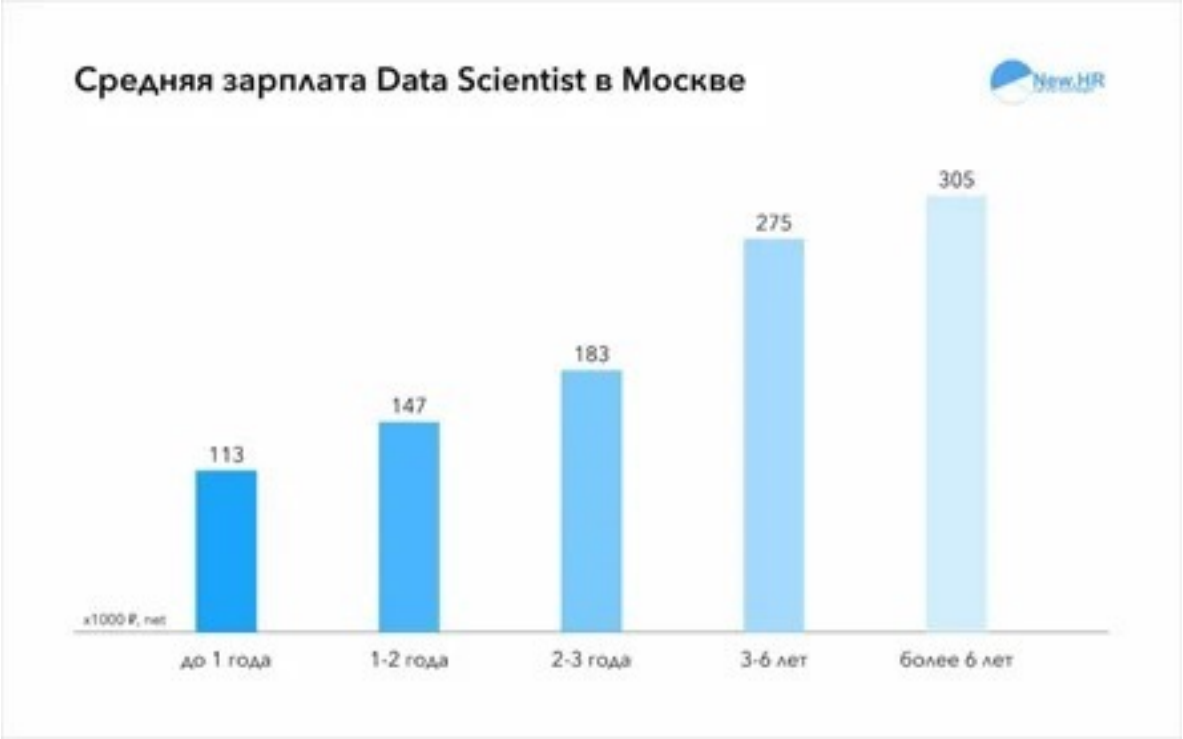
Анализ данных – это область математики и информатики, занимающаяся построением и исследованием наиболее общих математических методов и вычислительных алгоритмов извлечения знаний из экспериментальных (в широком смысле) данных.

Более популярный термин – **Data mining** (интеллектуальный анализ данных) – включает всевозможные методы классификации, моделирования и прогнозирования.

Специалистов, занимающихся анализом данных, называют **Data Scientist.**

ПОЧЕМУ ЭТО ИНТЕРЕСНО

Говорят, что **Data Scientist –** профессия будущего. Не менее часто говорят и о высоких зарплатах:



Знание языка Python – один из важнейших навыков Data Scientist’a.

БИБЛИОТЕКА PANDAS

pandas — программная библиотека на языке Python для обработки и анализа данных. Работа pandas с данными строится поверх библиотеки NumPy. Предоставляет специальные структуры данных и операции для манипулирования числовыми таблицами.

Название библиотеки происходит от термина «панельные данные» (**Pan**el **Da**ta), а не от милого медведя.

Человек	Год	Доход	Возраст	Пол
№ 1	2016	1300	27	1
№ 1	2017	1600	28	1
№ 1	2018	2000	29	1
№ 2	2016	2000	38	2
№ 2	2017	2300	39	2
№ 2	2018	2400	40	2

Это панельные данные

А это панда

УСТАНОВКА

Pandas включен в дистрибутив Anaconda, так что если вы устанавливали его себе на прошлом занятии, ничего делать не нужно, всё готово к работе.

Pandas можно установить с помощью менеджера пакетов pip:

```
pip install pandas
```

Если вы работаете в Jupyter Notebook, установку Pandas можно запустить прямо из ячейки кода:

`!pip install pandas` – команда воспринимается интерпретатором, как консольная, т.к. начинается со знака !

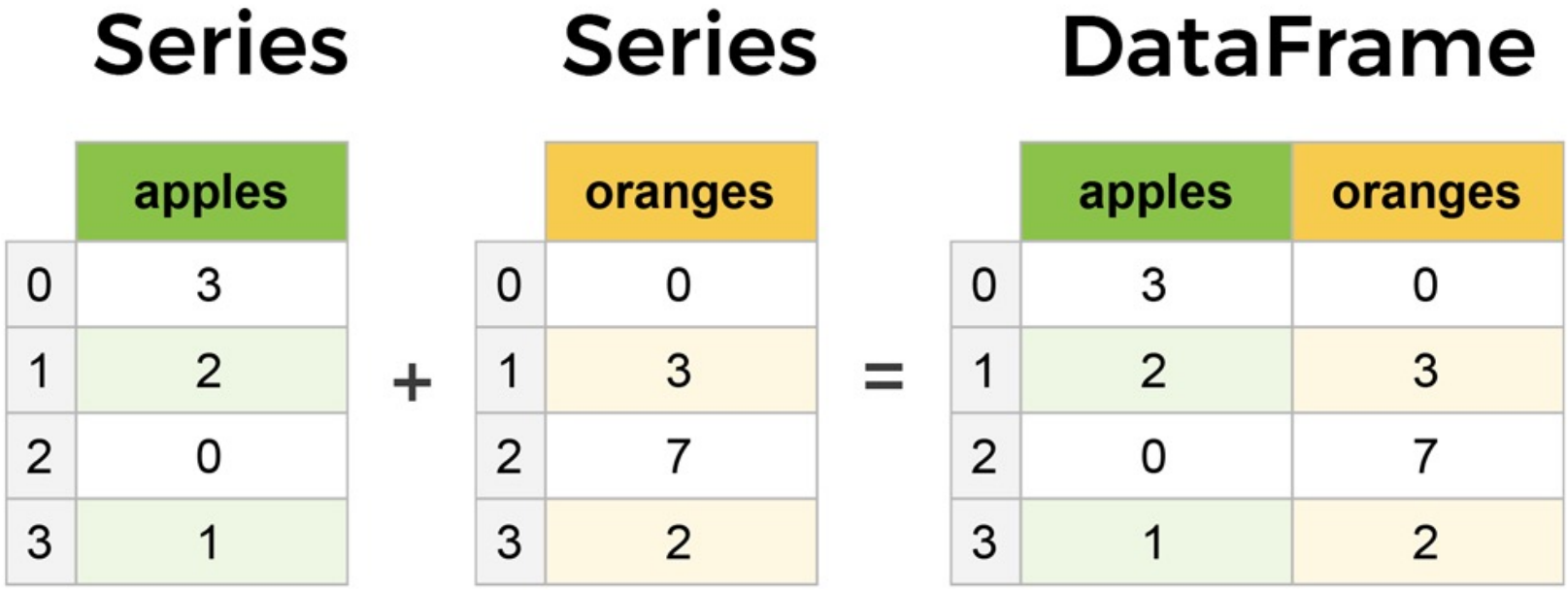
ОСНОВНЫЕ КОМПОНЕНТЫ

Самые главные структуры данных библиотеки: **DataFrame** и **Series**.

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

DataFrame — это проиндексированный многомерный массив значений.

Каждый столбец DataFrame является структурой Series



SERIES

В строковом представлении объекта Series, индекс находится слева, а сам элемент справа. Если индекс явно не задан, то pandas автоматически создаёт *RangeIndex* от 0 до N-1, где N общее количество элементов. У Series есть тип хранимых элементов, в нашем случае это int64, т.к. мы передали целочисленные значения.

```
import pandas as pd
```

```
series = pd.Series([5, 6, 7, 8, 9, 10])
print(series)
```

```
0     5
1     6
2     7
3     8
4     9
5    10
dtype: int64
```

Изучите блокнот Pandas, раздел Series

DATAFRAME

Объект **DataFrame** лучше всего представлять себе в виде обычной таблицы, ведь DataFrame является табличной структурой данных. В любой таблице всегда присутствуют строки и столбцы. Столбцами в объекте DataFrame выступают объекты Series, строки которых являются их непосредственными элементами. Jupyter поддерживает "красивое" отображение DataFrame'ов:

Country Code	country	population	square
kz	Kazakhstan	17.04	2724902
ru	Russia	143.50	17125191
by	Belarus	9.50	207600
ua	Ukraine	45.50	603628

Изучите блокнот Pandas, раздел DataFrame

ЗАПИСЬ И ЧТЕНИЕ ДАННЫХ

pandas поддерживает все самые популярные форматы хранения данных: csv, excel, sql, буфер обмена, html и многое другое. Чаще всего приходится работать с csv-файлами.

`to_csv('filename.csv')` – сохранить dataframe в csv-файл.

`read_csv('filename.csv')` – считать данные из файла filename.csv в dataframe.

Опционально передаётся аргумент sep, указывающий на используемый разделитель.

МЕТОДЫ РАБОТЫ С ДАННЫМИ

Для dataFrame можно выполнить следующие методы:

- head() – показать первые 5 записей фрейма
- tail() – показать последние 5 записей фрейма
- info() – получить общую информацию о фрейме: количество столбцов и колонок, их названия, типы, размер занимаемой памяти.
- describe() – рассчитать метрики по колонкам.
- groupby() – сгруппировать данные, в качестве аргумента передаётся колонка или массив колонок, по которым идёт группировка.
- plot() – построить график по dataFrame.

У перечисленных методов огромное количество аргументов для самых разных целей, про них читайте в документации:

https://pandas.pydata.org/pandas-docs/version/0.23.4/api.html#dataframe

Примеры использования смотрите в разделе Группировка и агрегирование данных блокнота Pandas.

СТАТИСТИЧЕСКИЙ АНАЛИЗ

Метрики, рассчитываемые методом describe:

Вывод	Значение
count	Подсчёт частоты того или иного события
mean	Среднее значение.
std	Стандартное отклонение (числовое значение, которое отображает изменение пределов данных).
min	Наименьшее число в наборе данных.
25%	25-й процентиль.
50%	50-й процентиль.
75%	75-й процентиль.
max	Максимальное число в наборе данных.

Метрики рассчитываются только для числовых колонок.

ДОМАШНЕЕ ЗАДАНИЕ

Проанализируйте данные о пассажирах Титаника из файла titanic.csv. Ответьте на следующие вопросы:

1. Какое количество мужчин и женщин ехало на корабле?

2. Какой части пассажиров удалось выжить? Посчитайте долю выживших пассажиров. Ответ приведите в процентах.

3. Какую долю пассажиры первого класса составляли среди всех пассажиров? Ответ приведите в процентах.

4. Какого возраста были пассажиры? Посчитайте среднее и медиану возраста пассажиров.

5. Какое самое популярное женское имя на корабле?

Отчёт оформите в виде Jupyter-блокнота.

СПАСИБО ЗА ВНИМАНИЕ !

ВОПРОСЫ ?

School of Computer Science