Тема 2. Библиотека Pandas: основные объекты данных, визуализация данных

Практические задания для самостоятельного выполнения

Задания выполняются по вариантам. Формулировки заданий общие для всех вариантов; конкретные условия, указанные в общей формулировке, выбираются в соответствии с номером своего варианта (для удобства представлены в отдельном файле).

Результаты выполнения заданий необходимо представить в виде двух файлов:

1) ноутбук в формате *ipynb*, содержащий программный код, результаты его выполнения, а также все необходимые пояснения, выводы и комментарии (в текстовых ячейках);
2) файл в формате *pdf* (или *html*), полученный путем экспорта (или вывода на печать) ноутбука из п. 1).

Внимание: в названии файлов должна обязательно присутствовать фамилия автора (например, 'Петров_02.ipynb'). Безымянные работы проверяться не будут.

Обратите внимание, что все необходимые для выполнения задания программные конструкции рассмотрены в учебных ноутбуках, размещенных в системе LMS. После изучения этих материалов выполнение задания не потребует больших усилий.

Максимальная оценка за выполнение задания вне аудитории – 1 балл. Дополнительные баллы (от 0 до 4) можно будет получить на следующем практическом занятии по результатам тестирования.

Внимание: самостоятельное и вдумчивое выполнение задания серьезно повышает вероятность успешного прохождения теста (будет проверяться понимание работы принципов работы с инструментарием и, в частности, умение понимать программный код).

Задача 1.

Средствами библиотеки Pandas создать серию или фрейм данных (в зависимости от варианта). В серии должно быть не менее 20 элементов, во фрейме не менее 15 строк и не менее двух колонок. Набор данных должен иметь индексы, колонки фрейма – осмысленные названия. Полученную серию (фрейм) вывести целиком, показать длину, размерность, число элементов. Покажите тип элементов. Подсчитайте число уникальных значений каждого столбца.

Задача 2.

На просторах интернета найти файл с расширением **csv**, представляющий собой датасет реальных (а не выдуманных) данных. Для поиска можно загуглить фразы типа «Датасеты для машинного обучения и анализа данных», «20 лучших датасетов». Запрещено брать датасеты «ирисы», «преступность в Чикаго», «пингвины», «недвижимость Бостона», «пассажиры Титаника», а также те, что рассматривались на парах. Датасет должен иметь не менее 100 строчек (записей) и не менее четырёх столбцов, как минимум два из которых числовые, один категориальный уникальный (без повторения значений, например, номера телефонов) и один категориальный с повторением значений (например, национальность).

Файл загрузить во фрейм **df** (переменную типа DataFrame). Если в исходном фрейме много колонок, то удалить лишние, оставив не более 6, удовлетворяющих условиям из предыдущего абзаца.

Вывести на экран список колонок, число строк фрейма, размерность, общее число элементов, проверить, есть ли пустые элементы. Вывести первые и последние три строки фрейма.

Для каждой колонки указать тип данных и число уникальных элементов. А для каждой числовой колонки дополнительно вывести значения максимального, минимального элементов и среднее арифметическое.

Задача 3.

По данным фрейма из второй задачи построить два наиболее показательных графика (диаграммы) разных типов. Обосновать, почему выбраны именно эти два типа. Графики (диаграммы) оформить максимально информативно.