БЛОК II. ФУНКЦИОНАЛЬНАЯ ЧАСТЬ

II.1 Файл «**tb-cases.txt**» содержит данные по числу выявленных случаев туберкулеза в США по отдельным штатам за 1993–2018 гг. Создать объект *DataFrame* с иерархическим индексом (*MultiIndex*), состоящим из названия штата и года.

Отобразить эти данные на наиболее подходящем графике и определить штат с самым высоким относительным ростом случаев заболевания туберкулезом за рассматриваемый период.

II.2 Численность населения каждого штата США за период 1993–2018 гг. со- держится в файле «**US-populations.txt**». Прочитать эти данные в объект *DataFrame* с созданием соответствующего индекса и проанализировать их с целью обнаружения любых интересных трендов. Затем объединить эти данные с данными из задачи **II.1**, чтобы определить штаты с наибольшим и наименьшим преобладанием случаев заболевания туберкулезом на душу населения в 2018 г.

II.3 Использовать метод **cut**(pandas) для классификации звезд в наборе данных из задачи **I.3** по их температуре с распределением звезд по группам «M», «K», «G», «F», «A», «B» и «O» с левыми границами (в K) 2400, 3700, 5200, 6000, 7500, 10 000 и 30 000.

Изменить исходный код решения этой задачи для отображения звезд на графике цветом, соответствующим их температуре, используя следующее отображение:

```
color_mapping = {
    'M': '#FFB56C', 'K': '#FFDAB5', 'G': '#FFEDE3',
    'F': '#F9F5FF', 'A': '#D5E0FF', 'B': '#A2C0FF',
    'O': '#92B5FF'
}
```

Совет: pandas предоставляет метод **map** для отображения значений, вводимых из существующего столбца, в значения, выводимые в новый столбец, с использованием словаря.

II.4 Выполнить повторный анализ данных из примера **П9.11** (данные в архиве – **books-examples-data.zip**, файл – **eg10-millikan-data.txt**), в котором рассматривался эксперимент Милликена с каплями масла, с использованием более точного приближения значения эффективной вязкости воздуха:

$$\eta = \frac{\eta_0}{1 + \frac{b}{ap}}$$

, где $p = 100.82$ кПа – давление воздуха, $\eta_0 = 1.859 \times 10^{-5}$ кг/м · с, $b = 7.88 \times 10^3$ Па · м, a – радиус капли.

II.5 Группа цифровых технологий Кембриджского университета (Cambridge University Digital Technology Group) ведет записи погодных условий, наблюдаемых с крыши собственного здания, начиная с 1995 г. Эти данные прдеставлены в файле **weather-raw.csv**, формат данных описан в **weather-raw-format.txt**[1].

Прочитать весь этот набор данных и выполнить его парсинг средствами библиотеки pandas, чтобы определить:

1. наиболее частое направление ветра;
2. самую высокую измеренную скорость ветра;
3. год с самым солнечным июнем;
4. день с наибольшим количеством осадков;
5. самую низкую измеренную температуру.

Следует отметить, что в этом наборе данных встречаются отсутствующие и некорректные точки данных.

II.6 Организация экономического сотрудничества и развития (ОЭСР) в рамках своей Международной программы по оценке образовательных достижений учащихся (Programme for International Student Assessment – PISA) публикует оценку образовательных систем по всему миру, проводя раз в три года тест, оценивающий функциональную грамотность школьников в разных странах мира и умение применять знания на практике. В тесте участвуют подростки в возрасте 15 лет. Оценка качества образования проводится по трем основным направлениям: грамотность чтения, математическая грамотность и естественно-научная грамотность.

Данные PISA в хронологической последовательности (по годам) содержатся в архиве **PISA-data.zip**. Прочитать эти данные в объект *DataFrame* (pandas) и использовать функциональные средства группирования для определения и визуализации:

- общей функциональной грамотности во всех исследуемых странах по времени;
- несоответствия между полами (если оно наблюдается) по каждому исследуемому направлению: чтение, математика и естественно-научная грамотность;
- корреляции между этими исследуемыми направлениями по всем странам.

[1]Оригинал даных можно найти по адресу: cl.cam.ac.uk/research/dtg/weather/

II.7 В файле **f1-data.csv** содержатся результаты каких-то сезонов гонок Формула 1 Гран-при. Составить рейтинг:

- пилотов по количеству побед на этапах;
- конструкторов по количеству побед на этапах;
- этапов (трасс) по усредненному самому быстрому кругу (по времени) в гонке.