Задание для домашней работы 1

В [таблице](https://docs.google.com/spreadsheets/d/1KjTBe7UVT6aF8I3MheO-8K-O4jfr9ippapJNWMN86fA/edit?usp=sharing) пары столбцов соответствуют протеомам шести бактерий и шести эукариотических организмов. Таблицы для всех известных белков данных организмов содержат имена белков согласно банку Uniprot и их дли́ны (количества аминокислотных остатков). Имя белка состоит из двух частей, вторая часть (после подчёркивания) называется “мнемоника организма”. Соответствие мнемоник организмам следующее:

**Бактерии:**

| Мнемоника | Организм |
| --- | --- |
| BACSU | *Bacillus subtilis* (сенная палочка) |
| CHLTR | *Chlamydia trachomatis* |
| ECOLI | *Escherichia coli* (кишечная палочка) |
| MYCMS | *Mycoplasma mycoides* |
| MYCTU | *Mycobacterium tuberculosis* (палочка Коха) |
| PROMA | *Prochlorococcus marinus* |

**Эукариоты:**

| Мнемоника | Организм |
| --- | --- |
| CAVPO | *Cavia porcellus* (морская свинка) |
| CHLVA | *Chlorella variabilis* (зелёная водоросль хлорелла) |
| DROVI | *Drosophila virilis* (плодовая мушка) |
| PHYIT | *Phytophthora infestans* (фитофтора) |
| PLABA | *Plasmodium berghei* (малярийный плазмодий) |
| YEAST | *Saccharomyces cerevisiae* (дрожжи) |

### Задание

1. Выбрать одну из шести бактерий и один эукариотический организм
2. Проиллюстрировать различия между распределениями длин белков двух выбранных организмов (бактерии и эукариота) тремя способами:
   1. Совместной гистограммой
   2. Парой “ящиков с усами” (box plots)
   3. Графиками эмпирических функций распределения (два графика на одном рисунке)
3. Написать отчёт, в который включить полученные изображения и комментарии к ним

Отчёты в формате pdf присылайте на мой адрес [sspirin@hse.ru](mailto:sspirin@hse.ru). Файл с отчётом лучше всего назвать своей фамилией латинскими буквами.

**Дедлайн — 23:59 10 декабря.**

### Требования к отчёту

Отчёт должен быть понятен постороннему читателю без обращения к тексту задания. То есть не должно быть просто картинок без объяснения, о чём это.

Обязательно описание методов (электронные таблицы, библиотеки языков программирования и т.п., выбор за вами), которыми получены изображения. Описание может быть сколь угодно кратким, но обязательно достаточным для повторного воспроизведения таких же изображений. В случае “ящиков с усами” необходимо объяснить, что в данном случае означают “усы” (поскольку существуют разные варианты).

Для хорошей оценки обязателен подбор шага гистограммы и других параметров такой, чтобы максимально информативно отобразить особенности двух распределений. Не возбраняется (если это покажется разумным) привести, например, не одну, а две совместные гистограммы с разными параметрами.

Каждое изображение должно быть прокомментировано: какие особенности двух наборов чисел можно, по вашему мнению, на нём увидеть.