Визуализация биомедицинских данных

Домашняя работа №2

Вам предстоит выполнить задания ниже в RMarkdown документе. После чего результат (не просто сам Rmd, но результат knit'a) загрузить в ваш GitHub репозиторий[1]. Домашнее задание сдаётся ссылкой на ваш репозиторий (проверьте настройки приватности).
Deadline: 12 ноября 2023 года

Домашнее задание оценивается по системе зачёт/незачёт. Зачёт ставится при выполнении любых 9 заданий. Любые спорные ситуации при оценке решаются в пользу студента.

Задания

1. Загрузите датасет life_expectancy_data.RDS (лежит в папке домашнего задания). Это данные с основными показателями, через которые высчитывается ожидаемая продолжительности жизни по метрике World Development Indicator на уровне стран[2]. В данных оставлены строки, относящиеся к положению женщин в 2019 г.
2. Сделайте интерактивный plotly график любых двух нумерических колонок. Раскрасть по колонке континента, на котором расположена страна[3].
3. Проведите тест, на сравнение распределений колонки `Life expectancy` между группами стран Африки и Америки. Вид статистического теста определите самостоятельно. Визуализируйте результат через библиотеку `rstatix`.
4. Сделайте новый датафрейм, в котором оставите все численные колонки кроме `Year`. Сделайте корреляционный анализ этих данных. Постройте два любых типа графиков для визуализации корреляций.
5. Постройте иерархическую кластеризацию на этом датафрейме.
6. Сделайте одновременный график heatmap и иерархической кластеризации. Содержательно интерпретируйте результат.
7. Проведите PCA анализ на этих данных. Проинтерпретируйте результат.
8. Постройте biplot график для PCA. Раскрасьте его по значениям континентов. Переведите его в `plotly`. Желательно, чтобы при наведении на точку, вы могли видеть название страны.
9. Дайте содержательную интерпретацию PCA анализу.
10. Сравните результаты отображения точек между алгоритмами PCA и UMAP.

---

[1] Есть два способа сделать это: первый лёгкий и не совсем корректный (но результат будет правильным), второй сложнее, зато поможет вам понять, как выстроить весь цикл работы в репозитории (детали хорошо объяснены в этом видео (спасибо Екатерине Фокиной за находку)). Во втором случае общая идея в том, что вы создаёте и клонируйте свой репозиторий, а потом настраивайте R, чтобы делать коммиты удобее).
[2] Источник: https://www.kaggle.com/datasets/kiranshahi/life-expectancy-dataset/data
[3] Plotly не всегда корректно ведёт себя во время knit – для него нужно настраивать .Rmd документ. Если вы столкнулись с тем, что у вас не-"нитится" из-за plotly – просто отмените выполнение чанка при сохранении кода в его настройках (eval=FALSE)

11. *Давайте самостоятельно увидим, что снижение размерности – это группа методов, славящаяся своей неустойчивостью.* Удалите 5 случайных колонок. Проведите PCA анализ. Повторите результат 3 раза. Наблюдаете ли вы изменения в куммулятивном проценте объяснённой вариации? В итоговом представлении данных на биплотах? С чем связаны изменения между тремя PCA?
12. *Давайте самостоятельно увидим, что снижение размерности – это группа методов, славящаяся своей неустойчивостью.* Создайте две дамми-колонки о том: (1) принадлежит ли страна к африканскому континенту, (2) Океании. Проведите PCA вместе с ними, постройте биплоты. Проинтерпрейтируйте результат. Объясните, почему добавление дамми-колонок не совсем корректно в случае PCA нашего типа.