Серебренников Д.

ИПП ЕУ СПб

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: 23:59 31 октября 2022 г.

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 insurance_cost.csv (лежит в папке домашнего задания). Это данные по базовым показателям здоровья индивида и сумме, которую страховая компания заплатила за его лечение в год. Обычно эти данные используют, чтобы потренироваться в предсказании того, как определённые характеристики индивида повышают траты страховой компании (и, соответственно, должны быть заложены в цену страховки).
2. Сделайте интерактивный plotly график отношения индекса массы тела и трат на страховку. Раскрасьте его по колонке smoker[2].
3. Сделайте тоже самое через ggplotly[3].
4. Кратко сделайте корреляционный анализ данных insurance_cost. Посмотрите документацию пакетов, которые мы проходили на занятии и, исходя из этого, постройте минимум два новых типа графика (которые мы не строили на занятии).
5. Превратите все номинативные переменные в бинарные/дамми. Т.е. sex и smoker должны стать бинарными (1/0), а каждое уникальное значение region – отдельной колонкой, где 1 говорит о наличии этого признака для наблюдения, а 0 – об

[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] 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)

[3] Сноска выше относится и к ggplotly

отсутствии[4]. Создайте новый датафрейм, где вы оставите только нумерические переменные.

6. Постройте иерархическую кластеризацию на этом датафрейме
7. (это задание засчитывается за два) Используя документацию или предложенный учебник[5] сделайте ещё несколько возможных графиков по иерархической кластеризации. Попробуйте раскрасить кластеры разными цветами[6].
8. Сделайте одновременный график heatmap и иерархической кластеризации
9. Проведите анализ данных полученных в задании 5 методом PCA. Кратко проинтерпретируйте полученные результаты.
10. В финале вы получите график PCA по наблюдениям и переменным. Сделайте кластеризацию[7] данных на нём по возрастным группам (создайте их сами на ваш вкус, но их количество должно быть не меньше 3).
11. (это задание засчитывается за два) Подумайте и создайте ещё две номинативные переменные, которые бы гипотетически могли хорошо разбить данные на кластеры[8]. Сделайте две соответствующие визуализации.
12. (это задание засчитывается за три) Давайте самостоятельно увидим, что снижение размерности – это группа методов, славящаяся своей неустойчивостью. Попробуйте самостоятельно поизменять дафрейм – удалить какие-либо переменные или создать их (создавайте только дамми переменные). Ваша задача – резко поднять качество вашего анализа PCA (при этом, фактически, оперируя всё теми же данными). Кратко опишите, почему добавление той или иной дамми-переменной так улучшает PCA.

[4] В работе с данными эта операция называется one hot-encoding
[5] С. 64-117
[6] С. 75
[7] В значении кода на строке 909 файла R_pro_work_with_graphics.Rmd
[8] В значении кода на строке 909 файла R_pro_work_with_graphics.Rmd