Домашнее задание

В файле `parfume.csv` приведены результаты маркетингового исследования, целью которого было выяснить предпочтения девушек от 18 до 25 лет относительно рекламы парфюмерной воды *Black & White* от *Carolina Herrera* (название изменено). Все девушки, участвовавшие в исследовании, были случайным образом распределены между контрольной и тестовой группой. Девушкам в контрольной группе показывали рекламный видеоролик в цветном варианте, а в тестовой — видеоролик в черно-белом варианте. Через некоторое время после демонстрации ролика фиксировалось, хотела бы девушка приобрести парфюмерную воду (значение 1) или нет (значение 0). По итогам анализа данных необходимо решить, какой из вариантов роликов оставить и использовать в дальнейшем.

Описание данных: `id` — id респондента, `age` — возраст респондента, `group` — группа (тестовая или контрольная), `vote` — решение о покупке (1 — хочет купить, 0 — не хочет, 2 — нет ответа).

Задание

Часть 1: R

1. Загрузите данные из файла `parfume.csv` в датафрейм `parf`.
2. Оставьте в датафрейме только те строки, которые соответствуют девушкам, предоставившим ответ на вопрос о решении купить/не купить парфюмерную воду. В последующих заданиях необходимо работать с обновленным датафреймом.
3. Для описания имеющихся данных постройте две столбиковые диаграммы (*bar plots*), которые показывают, сколько девушек хотели/не хотели бы приобрести парфюмерную воду *Black & White* в контрольной и тестовой группе (одна диаграмма для одной группы). Приведите графики в порядок, добавьте содержательные подписи и названия.

Подсказка: используйте функцию `table()` для получения числа нулей и единиц в столбце `vote`, а затем воспользуйтесь графической функцией `barplot()`.

4. Сохраните в вектор `test` ответы на вопрос о решении купить/не купить парфюмерную воду всех девушек, которые были в тестовой группе. Сохраните в вектор `control` ответы на вопрос о решении купить/не купить парфюмерную воду всех девушек, которые были в контрольной группе.

Подсказка: сначала отберите строки, соответствующие этим девушкам, а потом извлеките из полученного датафрема нужный столбец.

5. Сформулируйте нулевую гипотезу и одностороннюю альтернативную гипотезу, выдвигаемые для принятия решения о выборе одного из двух рекламных роликов (какой из роликов предпочла большая доля девушек). Считайте, что по мнению экспертов, ожидается, что черно-белый ролик будет более предпочтительным.
6. Используя бутстрэп с числом итераций $N = 1000$, найдите p-value для разности долей девушек, которые захотели купить *Black & White* в контрольной и тестовой группах. Для воспроизводимости результатов зафиксируйте стартовую точку с помощью `set.seed(123)`. Не забудьте о центрировании, которое обсуждалось в практическом задании 1.
7. Сделайте статистический вывод на основе полученного p-value, приняв уровень значимости равным 5% (отвергается или не отвергается нулевая гипотеза). Сделайте содержательный вывод о выборе рекламного ролика (какой из вариантов лучше).

Часть 2: Python

1. Загрузите данные из файла `parfume.csv` в датафрейм `parf` (используйте библиотеку `pandas`).
2. Оставьте в датафрейме только те строки, которые соответствуют девушкам, предоставившим ответ на вопрос о решении купить/не купить парфюмерную воду. В последующих заданиях необходимо работать с обновленным датафреймом.
3. Сохраните в датафрейм `test` строки, соответствующие девушкам из тестовой группы, а в датафрейм `control` — девушкам из контрольной группы.
4. Посчитайте на основе датафреймов `test` и `control` значения, необходимые для проверки гипотезы о равенстве долей с помощью функции `proportions_ztest`.
5. Проверьте гипотезу о равенстве долей с помощью функции `proportions_ztest`, учитывая, что альтернативная гипотеза является односторонней. Сделайте статистический вывод на основе полученного p-value, приняв уровень значимости равным 5% (отвергается или не отвергается нулевая гипотеза). Сделайте содержательный вывод о выборе рекламного ролика (какой из вариантов лучше).