Домашнее задание

Задание 1

1. Загрузите файл `housing.csv` с данными по ценам на квартиры в пригородах Бостона (одна строка – один пригород) и сохраните его в датафрейм. Описание показателей:

- `RM` – среднее число комнат;
- `LSTAT` – процент жителей низкого материального статуса:
- `MEDV` – медианное значение цены на дом в тысячах долларов.

   Посмотрите на датафрейм – откройте его в отдельном окне RStudio.

2. Выведите на экран первые и последние строки в датафрейме. Выведите информацию о типах столбцов в таблице («структуру» датафрейма). Проверьте, что все столбцы с числовыми данными считались как числовые, а не текстовые.

3. Выведите на экран сводные характеристики всех столбцов в датафрейме – описательные статистики. Наблюдаются ли серьезные отличия в медианных и средних значениях показателей (один из признаков наличия аномальных наблюдений или выбросов)? Насколько велик разброс значений цен на квартиры, если мы сравним минимальное значение и максимальное? Есть ли в каком-нибудь показателе с содержательной точки зрения нетипичные значения (отрицательные цены и прочее)?

4. Используя стандартные средства R (без `tidyverse`), добавьте в датафрейм столбец `MEDV_N`, который содержит медианные цены на квартиры из столбца `MEDV`, измеренные в долларах (не в тысячах долларов).

5. Используя стандартные средства R (без `tidyverse`), сохраните в датафрейм `small` только те строки, которые соответствуют пригородам, где медианная цена за квартиру (`MEDV`) больше 400 тысяч, но меньше 500 тысяч. Сколько таких пригородов?

   Подсказка: для определения числа строк в датафрейме можно воспользоваться функцией `nrow()`.

6. Используя средства библиотеки `tidyverse` (`dplyr`), создайте столбец `MEDV_LOG`, который содержит натуральные логарифмы значений медианных цен на квартиры из столбца `MEDV`.

7. Используя средства библиотеки `tidyverse` (`dplyr`), выведите на экран в отдельной вкладке строки, которые соответствуют пригородам, где процент населения, низкого по материальному статусу, составляет не менее 30%.

8. Используя стандартные средства R (без `ggplot2`), постройте гистограмму для показателя `LSTAT`. Добавьте заголовок и подписи к осям, измените цвет. Что можно сказать о распределении процента населения с низким материальным статусом? Какие значения преобладают, есть ли скошенность в сторону слишком больших или слишком маленьких значений? Сохраните полученный график в файл `hist.png`, используя код R (не кнопку *Export* в окне с графиком).

9. Выполните предыдущий пункт, но уже используя средства библиотеки `tidyverse` (графика с `ggplot2`).

Задание 2

1. Загрузите данные из файла `flats.csv` и сохраните в датафрейм. Посмотрите на датафрейм.

2. Сгруппируйте данные по показателю `brick` (дом из кирпича или нет). Определите, сколько домов каждого типа присутствует в данных. Каких домов больше?

3. Сгруппируйте данные по показателю `brick` (дом из кирпича или нет). Определите среднюю цену на квартиры (`price`) по каждой группе. Квартиры в каких домах, в среднем, дороже?

4. С помощью библиотеки `ggplot2` постройте гистограммы для цен на квартиры с разбиением на группы по показателю `walk` (находится ли дом в шаговой доступности от метро или нет). По группам – отдельное окно-фасетка для каждой группы в рамках одного графика. Распределение цен в какой группе менее скошенное (вправо или влево)?

5. С помощью библиотеки `ggplot2` постройте диаграммы рассеяния для показателей `totsp` (общая площадь квартиры в квадратных метрах) и `price` (цена квартиры), сделав цвет точек зависимым от показателя `walk`, а размер – от показателя `kitsp` (площадь кухни).

   Подсказка: с размером точек можно поступать точно так же, как с цветом, либо указывать внутри `geom_point()`, либо внутри `aes()` в зависисимости от постановки задачи. Аргумент для настройки размера точки – `size`.