Домашнее задание 4

Файл с выполненным заданием необходимо загрузить на Dropbox до дедлайна, указанного на сайте. Результат выполнения домашнего задания — файл с расширением `.Rmd`. Файл должен содержать код и ответы на вопросы.

Домашние задания, сданные после срока, оцениваются с использованием понижающих коэффициентов: опоздание в пределах часа – штраф 10% от полученной оценки, в пределах суток – штраф 20%, в пределах недели – штраф 50%. Домашние задания, сданные через неделю после указанного срока и позже, не принимаются и не оцениваются.

Если при проверке работ установлен факт нарушения академической этики, студент получает оценку «0» за данную работу. Работа студента, предоставившего свою работу для списывания, также аннулируется.

Файл `wgi_fh.csv` содержит показатели *Worldwide Governance Indicators* и индекс *Freedom House* за 2016 год. Описание переменных:

- `X` — id страны;
- `country` — название страны;
- `cnt_code` — код страны;
- `year` — год;
- `va` — индекс *Voice and Accountability*;
- `ps` — индекс *Political Stability and Lack of Violence*;
- `ge` — индекс *Government Effectiveness*;
- `rq` — индекс *Regulatory Quality*;
- `rl` — индекс *Rule of Law*;
- `cc` — индекс *Control of Corruption*;
- `fh` — индекс *Freedom Rating om Freedom House*.

Подготовка данных

Задание 1

Загрузите файл `wgi_fh.csv`, учитывая, что в качестве разделителя столбцов используется точка с запятой.

Задание 2

Посмотрите, каким образом в базе обозначены пропущенные значения. Замените все такие значения на «настоящие» пропущенные значения (`NA`).

Подсказка: чтобы заменить определённые значения сразу во всем датафрейме, можно поступить так — пример для замены значений 100 на 1000:

```
df[df == 100] <- 1000
```

Задание 3

Удалите в датафрейме все строки с пропущенными значениями (`NA`). Сохраните датафрейм в csv-файл `wgi_fh_new.csv`. Загрузите данные из файла `wgi_fh_new.csv` в R с учетом двух фактов:

- в качестве разделителя разрядов используется запятая (это нужно для того, чтобы R воспринимал записи вида "1,5" как числа, а не как текст);
- текстовые переменные из файла (*strings*) R должен считывать как текстовые, а не как факторные (*factor*).

Для выполнения этого пункта см. `?read.csv`.

Задание 4

Удалите пропущенные значения из загруженного датафрейма. Переименуйте столбец `X`, назовите его `id`.

Задание 5

Добавьте в датафрейм столбец `fh_type` со значениями `not free`, `partly free`, `free`. Для этого вам потребуется разбить все страны на группы по значению переменной `fh`. Соответствие значений индекса типу страны — см. в таблице 3 на стр. 18 в файле *Methodology* с сайта *Freedom House*. Преобразуйте тип полученного столбца в факторный.

Анализ данных

Задание 1

Определите, сколько стран разных типов (`free`, `partly free`, `not free`) в датафрейме. Каких стран больше всего?

Задание 2

Постройте график, который иллюстрировал бы ответ на первый вопрос (распределение стран по группам согласно классификации *Freedom House*). Приведите график в порядок: добавьте заголовок, скорректируйте оси и дайте им содержательные названия.

Задание 3

Постройте ящики с усами для индекса *Voice & Accountability* по каждой группе стран (`free`, `partly free`, `not free`) на одном графике (см. пример здесь). Есть ли в какой-нибудь группе стран нетипичные значения индекса *Voice & Accountability*? Если есть, укажите в какой. Обоснуйте свой ответ.

Задание 4

Поменяйте цвет графиков (покрасить все одним цветом или сделать три ящика разноцветными — на ваше усмотрение). Добавьте название графика.

Задание 5

Постройте гистограмму для индекса *Control of Corruption*. Поменяйте её цвет, добавьте заголовок и содержательные подписи к осям. Наложите на гистограмму график плотности нормального распределения с соответствующими параметрами. Как вы считаете, основываясь на графике, является ли распределение индекса *Control of Corruption* нормальным?

Задание 6

Проверьте формально, то есть используя статистический критерий, является ли распределение индекса *Control of Corruption* нормальным. Запишите нулевую гипотезу, название используемого статистического критерия. Проинтерпретируйте полученный результат: сделайте статистический и содержательный вывод. Примеры выводов см. здесь.

Задание 7

Постройте диаграмму рассеяния, которая иллюстрировала бы связь между показателями *Political Stability and Lack of Violence* и *Government Effectiveness*. Добавьте название графика и содержательные подписи к осям. Как вы считаете, основываясь на графике, есть ли связь между этими показателями? Если да, укажите направление связи (прямая или обратная) и силу связи (сильная, слабая, средняя).

Задание 8

Сделайте так, чтобы точки на диаграмме рассеяния из предыдущего пункта были полностью закрашены, а их цвета соответствовали типу государства по *Freedom House* (например, зелёные точки — *Free*, красным — *Not free*, синим — *Partly free*). Добавьте на график сокращённые названия стран (`cnt_code`) так, чтобы посмотрев на точку, можно было понять, какой стране она соответствует.

Подсказка: После строчки с `plot()` добавьте строчку с функцией `text()`. Пример:

```
x <- seq(-4, 4)
y <- x^2
labs <- LETTERS[1:9]
plot(x, y)
text(x, y, labels = labs, pos = 4)
```

Поправьте положение названий точек так, чтобы они не очень перекрывали друг друга (поэкспериментируйте с аргументами `pos` и `cex`, см. `help(text)`).

Задание 9

Проверьте формально, то есть используя статистический критерий, есть ли связь между показателями *Political Stability and Lack of Violence* и *Government Effectiveness*. Проинтерпретируйте полученный результат: сделайте статистический и содержательный вывод.

Задание 10

Используя функции из библиотеки `car`, постройте матрицу диаграмм рассеяния для всех индексов WGI (их 6). На главную диагональ матрицы поместите гистограммы для соотвествующих переменных. Не забудьте поправить в ячейках на главной диагонали названия переменных. Уберите с диаграмм рассеяния линию взвешенной регрессии (*loess regression*), оставьте только линию обычной линейной регрессии.

Задание 11

Постройте корреляционную матрицу для всех индексов WGI (их 6). Сохраните её в переменную `M`.

Задание 12

Установите библиотеку `corrplot`. Посмотрите документацию по работе с ней.

Задание 13

Используя уже созданную корреляционную матрицу `M`, постройте график для визуализации значений коэффициентов корреляции в ней. Считайте, что нам нужен график с эллипсом рассеяния. Есть ли среди пар показателей такие пары, коэффициент корреляции между которыми отрицательный? Обоснуйте свой ответ.

Задание 14

Постройте аналогичный график, состоящий из закрашенных клеток, где более темные клетки соответствуют более высокому коэффициенту корреляции (оттенки клеток для отрицательных и положительных коэффициентов корреляции отличаются). Такой график называется тепловой картой (*heatmap*).

Задание 15

Воспроизведите график.

Между какими показателями коэффициент корреляции наибольший? Наименьший? Есть ли различные пары показателей, коэффициент корреляции для которых одинаков?