Проект по курсу «Основы программирования в R»

Формат

1. Дедлайн: **3 июня 2020, 12:00**
2. Результат выполнения домашнего задания – четыре файла:
   - csv-файл с данными;
   - pdf-файл с описанием переменных в файле с данными (*codebook*);
   - Rmd-файл с кодом R, комментариями, графиками и прочим;
   - html-файл – результат компиляции («связывания») Rmd-файла из пункта выше.
3. Файлы, перечисленные выше, необходимо в указанный срок загрузить на Dropbox в виде zip-архива.

Задание

Знаком ⚠ (`danger`) обозначены обязательные требования к частям задания.

1. **Сформулировать вопрос, на который Вы хотите ответить в рамках данного мини-исследования.**

   Должен быть включен в Rmd-файл. Это может быть не один вопрос, а два-три связанных между собой вопроса или гипотезы. Подробную постановочную часть (проблема, исследовательский вопрос, цели, задачи, объект-предмет) писать не нужно.

2. **Выбрать данные для работы.**

   Файл с данными может быть любой. Исходный формат данных может быть любой (`xlsx`, `dta`, `sav`, `txt` и проч.), но для выполнения этого задания файл нужно сохранить в формате `csv`.

   ⚠ Датафрейм не должен быть совсем «чистым»: пусть в нем будут лишние показатели, которые Вы потом выкинете, пропущенные значения, не интересующие Вас страны/регионы/респонденты, что-то, что при подготовке к работе нужно будет убрать/изменить/преобразовать).

3. **Загрузить csv-файл с данными в R. Подготовить данные для дальнейшей работы.**

   Убрать лишние/добавить недостающие переменные, отфильтровать наблюдения, поменять типы переменных, переименовать столбцы или строки датафрейма и.т.д.

   ⚠ Подготовка данных должна быть выполнена средствами библиотеки `tidyverse`.

   ⚠ Этап подготовки данных должен быть отражен в Rmd-файле (код и описание словами, что делается и зачем).

4. **Создать *codebook* – файл с описанием переменных в данных.**

   Здесь и далее имеется в виду датафрейм с учетом преобразований из пункта 3.

   ⚠ Формат файла – pdf (обычный pdf, нет необходимости создавать его в RStudio или LaTeX).

⚠ *Codebook* должен содержать указание на источник данных (при необходимости ссылки), названия переменных (как они названы в датафрейме), описание переменных (что за показатели), типы переменных (шкалы), пояснения к значениям (единицы измерения, сокращения, закодированные значения). См. минимальный хороший *codebook* на примере базы данных по плебисциту в Чили.

5. **Подготовить описание данных – код и выдачи R.**

⚠ Описание датафрейма должно включать ответы на следующие вопросы:

1) Сколько в датафрейме наблюдений и переменных? 2) Какие это переменные, какого типа? 3) Есть ли в датафрейме пропущенные значения? Если да, то сколько? Наблюдаются ли какие-нибудь паттерны пропущенных значений? Какие?

⚠ Описание данных должно содержать описательные статистики для всех переменных в датафрейме. Для переменных интереса должны быть построены графики, отражающие распределение данных (столбчатые/круговые диаграммы, ящики с усами, скрипичные диаграммы, гистограммы и прочие). Должно быть не менее 3 графиков, из них как минимум 2 должны быть построены с помощью библиотеки `ggplot2`.

⚠ Этап описания данных должен быть отражен в Rmd-файле (код и описание словами, что делается и зачем).

6. **Провести анализ данных.**

Этот этап полностью зависит от целей исследования. Сюда может входить исследование формы распределения данных (например, проверка нормальности), сравнение средних значений/распределений (критерий Стьюдента, ANOVA, критерий Уилкоксона, критерий Краскела-Уоллиса), выявление связей между переменными (таблицы сопряженности, критерий хи-квадрат, корреляционные матрицы, коэффициенты корреляции и их значимость и прочее), построение регрессионных моделей.

⚠ Должно быть не менее 3 графиков, построенных с помощью `ggplot2` или других библиотек (не базовыми средствами R типа `plot`).

⚠ Должны быть использованы различные статистические критерии, проверки статистической значимости (не менее 2).

⚠ Этап анализа данных должен быть отражен в Rmd-файле (код и описание словами, что делается и зачем).

7. **Проинтерпретировать результаты.**

Содержательные выводы на основе результатов, полученных в пункте 6. Должны быть включены в Rmd-файл.