Большое домашнее задание по курсу «Основы программирования в R»

Большое домашнее задание весит 40% от накопленной оценки (не входит в оценку за самостоятельную работу наряду с обычными текущими домашними заданиями).

Сроки

Дедлайн: 16 марта 2018, 22:00

Предварительный дедлайн: 26 февраля 2018, 22:00

Предварительный дедлайн установлен для тех студентов, которые хотят сдать полностью выполненное задание или любую его часть заранее, чтобы получить фидбек, устранить недочеты и к основному дедлайну загрузить отредактированный вариант. Это дополнительная опция, по желанию. По заданиям, высланным после 26 февраля получить развернутый фидбек уже не получится (но возможность задавать вопросы никуда не исчезает).

Формат

Результат выполнения домашнего задания – четыре файла:

- csv-файл с базой данных
- pdf-файл с описанием переменных в базе данных (codebook)
- Rmd-файл с кодом R, комментариями, графиками и проч.
- html-файл – результат компиляции («связывания») Rmd-файла из третьего пункта

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 базу данных для работы.**

   База данных может быть любой. Исходный формат базы данных может быть любой (`xlsx`, `dta`, `sav`, `txt` и проч.), но для выполнения этого домашнего задания ее нужно сохранить в формате `csv`.

   ⚠ База данных не должна быть совсем «чистой»: пусть в ней будут лишние показатели, которые Вы потом выкинете, пропущенные значения, не интересующие Вас страны/регионы/респонденты, что-то, что при подготовке базы к работе нужно будет убрать/изменить/преобразовать).

3. **Загрузить csv-файл с базой данных в R. Подготовить базу данных для дальнейшей работы.**

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

⚠ Подготовка базы данных должна быть выполнена средствами библиотеки `dplyr`.

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

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

   Здесь и далее имеется в виду база данных 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 базе), 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 (семинар 27 сентября).

5. **Подготовить описание базы данных (не codebook, код и выдачи R).**

⚠ Описание базы данных должно включать ответы на следующие вопросы:

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

⚠ Описание базы данных должно содержать описательные статистики для всех переменных в базе. 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. **Интерпретация результатов.**

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