

173 группа: выслать задание до 13:30 17.03.2018 на почту psdataculture@gmail.com

174 группа: выслать задание до 16:30 17.03.2018 на почту psdataculture@gmail.com

172 группа: выслать задание до 12:00 12.03.2018 на почту mbburova@gmail.com

171 группа: выслать задание до 12:00 13.03.2018 на почту mbburova@gmail.com

Задание

У вас на руках выгрузка данных из Базы World Bank Open Data.

Мы скачали данные, которые в нашем представлении могли бы объяснить потенциальную длительность жизни в различных странах:
<http://databank.worldbank.org/data/reports.aspx?source=world-development-indicators#>

В данных вы найдете параметр 2015 [YR2015] - Life expectancy at birth, total (years) [SP.DYN.LE00.IN] (далее по тексту - Life expectancy) и ряд других показателей, описание которых дано на другом листе, по всем странам, доступным из World Data Bank.

Ваша задача: выяснить, какие из предложенных факторов оказывают значимое влияние на длительность жизни и могут служить хорошим предиктором для моделей предсказания длительности жизни.

Формат сдачи:

- 1) Фамилия.xlsx - excel-файл с вашими расчетами
- 2) Фамилия.docx - word-файл с вашими итогами

1. Нужно очистить данные от пропущенных значений

Подсказка: нужно воспользоваться макросом с предыдущего занятия.

2. На новом листе “Визуальный анализ” постройте гистограмму и частотную таблицу по данным переменной Life expectancy

Подсказка: округлите данные до целых функцией ОКРУГЛ, воспользуйтесь опцией “Гистограмма” из надстройки “Пакет анализа данных”. Сделайте так, чтобы было 11 карманов в интервале от 51 до 84 (т.е. вам вручную надо задать карманы 51-54, 54-57 и т.д.)

3. Далее необходимо отобрать факторы, которые могут оказывать влияние на life expectancy.

Подсказка: вам нужно провести корреляционный анализ из надстройки “Пакет анализа данных” на новом листе “Correlation Analysis”. Если коэффициент корреляции между факторами меньше по модулю 0.75, переменные нужно включать в анализ. Если связь между факторами больше 0.75 по модулю, то вы должны осмысленно выбрать переменную, которую включите далее в анализ (т.е. одна будет включена в модель, вторая нет - не забудьте прокомментировать, что убрали и что оставили).

Выделите с помощью условного форматирования сильно связанные факторы и сделайте вывод. (разумеется, life expectancy не нужно включать в анализ)

4. Постройте регрессионную модель с графиками на уровне надежности 95%.
5. Сделайте выводы:
 - a. насколько хорошая регрессия получилась (и подтверждение/опровержение этого факта)
 - b. какие факторы из включенных вами в статистический анализ оказались значимы
6. Оформите ваши рассуждения в виде .doc файла, в котором опишите:
 - a. Какую задачу вы решали
 - b. Приведите таблицу и гистограмму из п.1
 - c. Приведите таблицу с корреляциями и прокомментируйте наблюдаемую картину, а также прокомментируйте, если вы отобрали только некоторые переменные (или не отобрали, то почему)
 - d. Приведите таблицу с регрессией (не всю): коэффициенты, р-значения
 - e. Отрадите комментарии п.5 в тексте