**Сбор и анализ данных**

1. [Соскрапить данные с любого сайта](https://medium.com/nuances-of-programming/%D1%80%D1%83%D0%BA%D0%BE%D0%B2%D0%BE%D0%B4%D1%81%D1%82%D0%B2%D0%BE-%D0%BF%D0%BE-%D0%B2%D1%8D%D0%B1-%D1%81%D0%BA%D1%80%D0%B5%D0%BF%D0%B8%D0%BD%D0%B3%D1%83-%D0%BD%D0%B0-python-%D0%BF%D1%80%D0%B8%D1%91%D0%BC%D1%8B-%D0%B8-%D1%85%D0%B8%D1%82%D1%80%D0%BE%D1%81%D1%82%D0%B8-71a24a678c92) ([еще](https://m.habr.com/ru/company/plarium/blog/460675/))
2. Выполнить оптимизацию структуры данных
3. [Очистить данные и подготовить к дальнейшему анализу:](http://socialdata.org.ua/manual3/)

* удалить лишние колонки;
* колонки должны быть названы латиницей;
* привести в соответствие типы данных в колонках;
* выявить отсутствующие значения. если объем собранных данных позволяет (несколько тысяч записей и более), удалить записи с отсутствующими значениями, в противном случае заменить их средними значениями.

4. Определить виды шкал и переменных

5. Создать файл метаданных полученного датасета

6. Проанализировать меры центральной тенденции для каждого признака. Какие из них лучше подходят? Почему?

7. Построить гистограмму распределения признака (любого) и определить тип распределения

8. Оценить плотность вероятности признака ядерной оценкой плотности

9. Вывести график плотности вероятности

10. Вывести описательные статистики. Проанализировать выборку

11. Визуализировать описательные статистики на графике BoxPlot. дать объяснение

12. Есть ли в данных аномалии, выбросы? Что с ними делать?

13. Определить есть ли корреляция между признаками:

* построить графики рассеяния
* визуализировать регрессионную модель, используя seaborn.lmplot

#### **Поради щодо охайно структурованих даних (від Карла Бромана):**

1. Будьте послідовними.

* Використовуйте послідовні коди для категоріальних змінних.
* Використовуйте один фіксований код для відсутніх значень.
* Використовуйте послідовні назви змінних.
* Ідентифікатори об’єктів мають бути записані відповідно до однієї системи.
* Використовуйте однакову структуру даних в різних файлах.
* Називайте файли за однією системою, і використовуйте її послідовно.
* Використовуйте один загальний формат для всіх дат, переважно РРРР-ММ-ДД, наприклад 2016-06-21.
* Будьте уважними щодо зайвих пробілів всередині комірок (клітинок в таблиці). Порожня клітинка (дані відсутні) відрізняється від начебто порожньої клітинки з пробілами. Тому «Київ» машина сприйматиме інакше, ніж « Київ» (з пробілом), і дані потребуватимуть додаткової очистки.

1. Найкращий спосіб збереження даних в таблиці — коли у колонках змінні, а в рядках — об’єкти (спостереження). У першому рядку мають міститися назви змінних (не використовуйте більше одного рядка для назв змінних).
2. Заповнюйте всі комірки в таблиці, навіть коли дані відсутні — в такому разі варто мати спеціальне кодування для відсутніх даних (nan).
3. Вставляйте лише одне значення в комірку. Наприклад, у комірку може бути записана вага, наприклад «45 кг». Краще написати просто «45», а одиниці виміру винести в назву колонки, наприклад «weight\_kg». А ще краще, назвати колонку «weight», а одиниці виміру винести в окремий словник даних.
4. Створіть словник даних — окрему таблицю, в якій пояснювалися би змінні. Це буде частиною метаданих — дані про дані. Такий словник може містити:
   * точну назву змінної, як це вказано в наборі даних («технічну» назву змінної);
   * більш читабельну назву змінної, що може використовуватися, наприклад, для візуалізації даних;
   * більш розлогий опис змінної;
   * одиниці виміру;
   * очікувані максимальні та мінімальні значення змінних, наприклад.
5. Первинні дані не мають містити в комірках таблиці підрахунків чи формул.
6. Не використовуйте виділення кольором чи типом шрифта як дані. Наприклад, у якійсь таблиці в Excel-і можна виділити рядки різним кольором, щоб позначити окремо чоловіків і жінок. Так робити не треба, краще створити окрему змінну (колонку), і відповідно закодувати кожен рядок.
7. Називайте речі (змінні, файли) зрозуміло. Це може бути важко, але краще добре подумати, щоб назва була влучна і містка. Як правило, не варто використовувати пробіли в назвах змінних (тобто в заголовках колонок у таблицях з даними) або назвах файлів, замість них використовуйте дефіси (-) чи підкреслення (\_).
8. Регулярно робіть резервні копії ваших даних. І зберігайте їх в різних місцях.
9. Застосовуйте функцію перевірки даних до клітинок (наприклад в Excel-і — [goo.gl/TFB7vC](http://goo.gl/TFB7vC) ) для уникнення помилок

Джерело: *[kbroman.org/dataorg/](http://kbroman.org/dataorg/)*

Зберігайте дані у простих текстових форматах (таких як CSV). Використовуючи кирилицю, чи інші символи, що виходять за межі ASCII («базової латинки»), будьте дуже уважні щодо кодування, в якому зберігаєте файл з даними. Всім буде спокійніше, якщо це UTF-8.