# Чистка и предобработка данных

Вы научитесь:

* работать с библиотекой Pandas для очистки данных
* обрабатывать типичные проблемы данных в датасетах:
* наличие дубликатов, одинаковых записей
* нарушение уникальности
* противоречивость записей
* отсутствующие (или пропущенные) значения
* недопустимые значения или форматы
* орфографические ошибки и опечатки
* многозначность, перестановка слов
* аномальные значения, выбросы
* вложенные структуры данных

**Задание**

Маркетинговый отдел компании, занимающейся развитием киберспорта, планирует запустить таргетированную рекламу среди пользователей «ВКонтакте». В связи с этим им необходимо изучить целевую аудиторию и составить портреты пользователей, увлекающихся киберспортом, в том числе в разрезе онлайн-игр. Нашему коллеге-маркетологу необходимо построить дашборд с анализом выборки пользователей из онлайн-магазина, расположенного на платформе «ВКонтакте». Специалист по сбору данных представил датасет с выгруженной информацией по профилям в формате CSV. Однако данные в нем в том виде, в котором они есть, не подходят для построения дашборда.

Наша задача — очистить данные и привести их в нужный формат.

Какая информация важна для нашего коллеги-маркетолога:

1. id пользователя
2. возраст (кол-во лет)
3. пол (1: женский, 2: мужской, 0: не указан)
4. город проживания (если указан)
5. страна проживания (если указана)
6. количество подписчиков (если указано)
7. дата последнего входа в магазин
8. сумма трат в магазине (значение обязательно должно быть указано, пользователей с неуказанным значением необходимо исключить; значения должны быть соизмеримы, т.е. указаны в единой валюте)
9. наименования игр, которыми интересуется пользователь (значение обязательно должно быть указано, пользователей с неуказанным значением необходимо исключить)

Формат файла с результатами: xlsx или csv

Ссылка на блокнот Colab:

**Тема 1. Введение: Cleansing, Cleaning или Scrubbing**

Какие англоязычные термины называют процесс очистки данных?

* Data С: **Cleaning** – легкая очистка данных
* Data С: **Cleansing –** глубокая очистка данных
* Data S: **Scrubbing –** максимальная очистка данных

**Тема 2. Проблемы данных и процесс их очистки**

1. Что такое датасет?

Необработанные данные которые мы получили

2. Какие типичные проблемы данных в датасетах?

* Недопустимые значения
* Перестановка слов
* Вложенные структуры данных
* Аномальные значения
* Опечатки
* Отсутствующие значения

3. В чём важность этапа очистки данных?

В соответствии с принципом GIGO (мусор получаем – мусор отдаем) – от качества данных зависит конечный результат

4. Какие выделяют показатели качества данных в датасетах?

1. Согласованность - Непротиворечивость данных в различных связанных наборах данных
2. Точность - Необходимый уровень детализации данных

1. Достоверность - Корректность описания объекта

1. Интерпретируемость - Возможность понимания и извлечения смысла из данных

1. Полнота - Достаточность заполнения наборов данных и их атрибутов

1. Актуальность - Степень соответствия данных на определенный момент времени

5. Какие этапы включает процесс очистки данных?

1. Оценка качества данных
2. Проверка
3. Очистка
4. Повторная проверка
5. Отчетность

**Тема 3. Библиотека Pandas: подгрузка и обзор датасета**

1. Почему библиотека Pandas — главный инструмент аналитика?

Большой арсенал инструментов для обработки данных, ее возможности возможности:

* Чтение файлов различных форматов
* Манипулирование массивами данных
* Совмещение массивов данных
* Очистка данных
* Переформатирование данных
* Срезы выборки данных
* Группировки данных
* Создание сводных таблиц
* Работа с временными рядами
* Статистические функции
* Визуализация
* прочее

2. Какие объекты используются в Pandas?

DataFrame – прямоугольная таблица значений(состоит из Series) значение «индекс» обязательно. Строка – пример, колонка – признак.

3. Какие типы данных используются в Pandas?

Series – Объекты DataFrame

* Object – str or compilation
* Int64 – int
* Float64 – float
* Bool – bool
* Datetime – datetime
* Timedelta – разность между datetime
* Category - Ограниченный список текстовых значений

4. Как быстро создать объект DataFrame?

Метод .DataFrame – создает реляционную таблицу.

Df = pd.DataFrame(data)

5. Какие форматы файлов хранят табличные данные?

Csv – текст с разделителем

Xlsx - привычная таблица

Выявление проблем:

* Наличие признаков не представляющих ценности в рамках кейса
* Возможно наличие дубликатов
* Наличие незаполненных значений
* Наличие вложенных значений
* Приведение признаков в необходимый формат
* Необходимость расчет признака (возраст)
* Наличие закодированных значений (пол)
* Другие проблемы, которые могут быть выявлены в процессе предобработки данных

**Тема 4. Исключение признаков, работа с дубликатами**

1. Признак — это набор данных, который описывает в изучаемое в задаче явление.

Df.drop(colums=colums) – удаление столбцов

Не все признаки могут быть полезны, поэтому на этапе обработки мы отбираем необходимые.

Мультиколониарность → статистика

Отбор признаков (feature selection) → машинное обучение

1. Дубликат — это одинаковые записи или наблюдения попавшие в один датасет.

Полные дубликаты – исключаются из датасета.

Реализация поиска и исключения дубликатов в pandas :

Df.duplicated() – поиск дубликатов, передаем параметр subset. Т.е.: Df.duplicated(subset=[‘bdate’, ‘sex’])

Df.duplicated().sum() – суммирует все дубликаты

Df.drop\_duplicates() – удаляет все дубликаты, туда так же добавляется сабсет.

3. Как обрабатывать полные дубликаты?

удалять

4. Топ-3 вывода:

1.Надо внимательнее смотреть темы без теста, там может быть что-то полезное.

2. Не все то золото, что блестит.

**Тема 5. Работа с пропущенными значениями**

1. Пропуск — это отсутствие значения в данных.

2. Какие бывают пропуски в данных?

По характеру возникновения :

* Реальные – возникли в результате отсутствия ответа на поставленный вопрос, данные не записались и прочее
* Искусственные – возникают на этапе обработки, когда возникают заранее нереалистичные значения

По степени случайности:

* Полностью случайные missing complexly at random – нет данных вообще
* Случайные missing at random – когда ответ зависит от другого факта/фактора
* Неслучайные систематические not missing at random

Nan – тип float, с ним можно проводить математические операции в отличии от None – который имеет NoneType

1. Как проверить данные на пропущенные значения?

Df.isna() – проверка на пропуски

Df.isnull() - проверка на пропуски

Df.notna() – проверка на заполнение

Df.notnull() - проверка на заполнение

Df.info() – сводный обзор датафрецма

Df.isna().sum() – количество пропусков

Df.isnull().mean()\*100- процент пропусков

4. Что делать с пропусками?

Исключить:

Df.dropna() – построчное исключение

Попарное удаление – удаляются только обьекты у которых неизвестно 1-2 переменных

Восстановить:

Df.[‘city’].fillna(‘Moskow’) – статистические методы пример восстановления центральной тенденции химерой

Сохранить:

Методы анализа на которые не влияют отсутствующие значения

5. Как обрабатывать неслучайные | систематические пропуски?

**Тема 6. Работа с JSON и вложенными структурами**

1. JSON-формат — это …

2. Что может быть использовано в качестве значений в JSON?

1. : …
2. : …
3. : …
4. : …
5. : …

3. Как работать с JSON в Python и Pandas?

4. Что делает функция eval()?

5. Для чего используются строковые методы replace() и split()?

**Тема 7. Преобразование данных в необходимый формат**

1. Когда применяется модуль datetime?

2. Метод astype() — это …

3. Как преобразовать данные в формат целого числа?

4. Как преобразовать данные в формат числа с плавающей точкой?

**Тема 8. Расчет дополнительных признаков. Кодирование и декодирование**

1. Как получить текущие дату и время в Pandas?

2. Кодирование информации — это …

3. Декодирование информации — это …

**Тема 9. Форматирование текстовых данных. Проверка и сохранение датасета**

1. Какие методы Pandas используются для работы с текстовыми данными?

2. Метод explode() — это …?

3. Как проверить результаты очистки данных?

4. Как сохранить датасет в csv | excel?