# Практическая работа №19

**Тема практического занятия:** Обработка отсутствующих данных

**Цель практического занятия:** Овладеть навыками обработки отсутствующих данных в различных форматах и сохранение обработанных данных.

**Уметь:**

- Загружать данные из различных форматов (CSV, Excel, JSON) в pandas DataFrame.

- Обрабатывать отсутствующие данные, используя различные методы.

- Визуализировать данные до и после обработки.

- Сохранять обработанные данные в новые файлы.

**Знать:**

- Как работают основные методы обработки отсутствующих данных, такие как заполнение средними значениями.

Перечень оборудования, необходимого для выполнения задания: компьютер, монитор, компьютерная мышь, клавиатура

**Общие теоретические сведения:**

Отсутствующие данные (или NaN) могут оказаться проблемой при анализе данных. Обработка отсутствующих данных включает в себя заполнение пропущенных значений так, чтобы они не влияли на результаты анализа.

В pandas, для обработки отсутствующих данных, часто используются следующие методы:

fillna(): Заполняет отсутствующие значения определенным значением, например, средним или медианой.

CSV (Comma-Separated Values): Текстовый формат, где значения разделены запятыми. В этом формате данные были сохранены и загружены из файла 'data.csv'.

Excel: Формат табличных данных, используемый в программе Microsoft Excel. Данные были сохранены и загружены из файла 'data.xlsx'.

JSON (JavaScript Object Notation): Формат обмена данными, представляющий собой текст в формате ключ-значение. Данные были сохранены и загружены из файла 'data.json'.

pandas: Используется для работы с табличными данными, загрузки и сохранения данных.

numpy: Предоставляет функциональность для работы с числовыми массивами и матрицами.

matplotlib.pyplot: Библиотека для создания графиков и визуализации данных.

seaborn: Надстройка над matplotlib, предоставляющая более высокоуровневый интерфейс для визуализации статистических данных.

Обработка отсутствующих данных важна по нескольким причинам:

Исключение искажений: Присутствие отсутствующих данных может исказить результаты анализа, особенно при вычислении статистик.

Предотвращение ошибок: Некоторые алгоритмы и функции не могут обрабатывать отсутствующие значения, что может привести к ошибкам выполнения.

Улучшение качества моделей: В машинном обучении обработка отсутствующих данных может улучшить производительность моделей, обученных на этих данных.

Сохранение целостности данных: Обработка отсутствующих данных позволяет сохранить целостность набора данных и избежать искажений при анализе.

**Задание:**

1. Загрузите данные из созданных файлов: CSV, Excel, JSON.

2. Обработайте отсутствующие данные, используя средние значения по соответствующим столбцам.

3. Визуализируйте распределение возраста до и после обработки для каждого формата данных.

4. Сохраните обработанные данные в новые файлы.

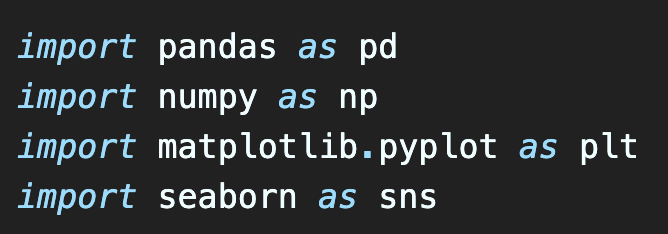
Указания по технике безопасности:

1. Регулярно сохраняйте свою работу, чтобы избежать потери данных.

Технология выполнения работы (этапы, последовательность действий):

**Ход выполнения работы:**

1. Импорт данных

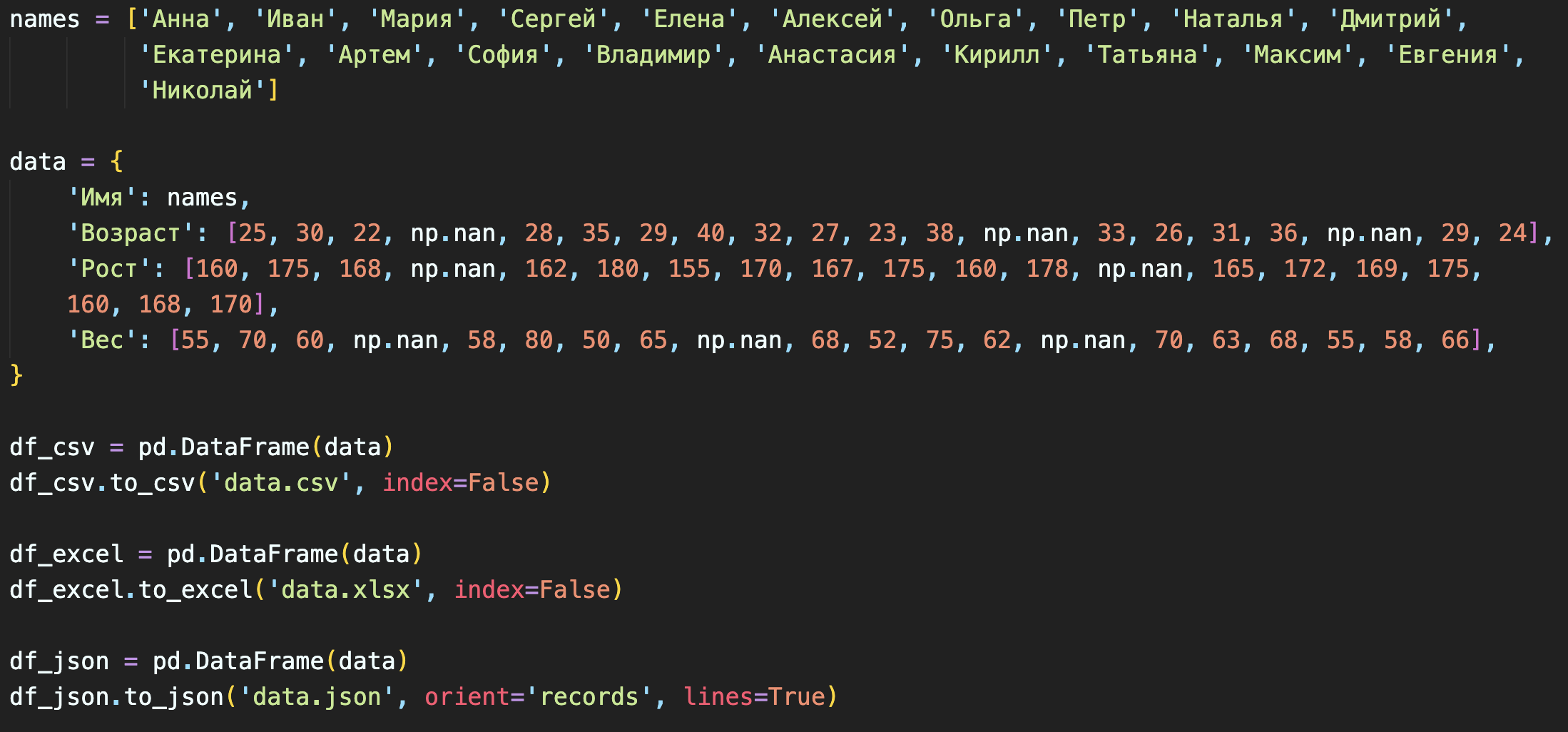


1. Создание и сохранение данных в различные форматы

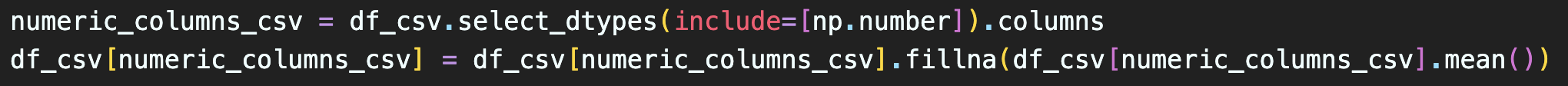


1. Создание файлов csv, xlsx и json

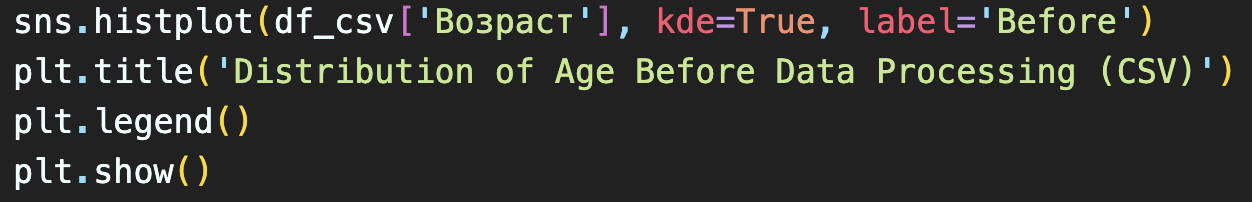
Переменная names служит для разделите имена на отдельные строки.



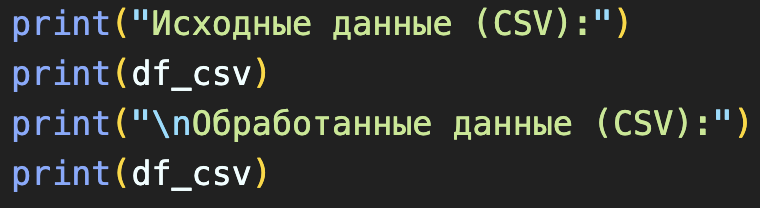
1. Заполнение пропущенных значений средними (исключаем столбец 'Имя')



1. Визуализация данных до обработки



1. Сравнение результатов



1. Сохранение обработанных данных в новый CSV-файл



1. Самостоятельная работа

Согласно пункту 5-7 выполните самостоятельно для файлов xlsx и json.

Заполнение пропущенных значений средними

Визуализация данных до обработки

Сравнение результатов

Сохранение обработанных данных в новый Excel-файл

Заполнение пропущенных значений средними

Визуализация данных до обработки

Сравнение результатов

Сохранение обработанных данных в новый JSON-файл

При сохранении файла json добавьте атрибут orient='records' и , lines=True и исключите index = его быть не должно.

**Требования к отчету:**

1. В отчете должны быть представлены графики распределения возраста до и после обработки для каждого формата данных.

2. Приведите код и объяснение шагов обработки данных.

3. Укажите, какие методы использовались для обработки отсутствующих данных.

4. Предоставьте созданные файлы с обработанными данными.

**Контрольные вопросы:**

1. Какие методы используются для обработки отсутствующих данных в pandas?

2. Какие форматы данных вы использовали в данной практической работе?

3. Какие библиотеки были задействованы в процессе выполнения работы?

4. Какие этапы выполнения работы представлены в коде?

5. Почему важно обрабатывать отсутствующие данные перед анализом данных?