

САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

Факультет прикладной математики — процессы управления

Программа бакалавриата

«Большие данные и распределенная цифровая платформа»

ОТЧЁТ

по лабораторной работе №4

по дисциплине «Восстановление пропусков»

Студент гр. 22Б16-пу

Преподаватель

Шарабарин М.С.

Дик А.Г.

Санкт-Петербург

2025 г.

Содержание

1 Цель работы	3
2 Использованные методы импутации:	3
3 Описание задачи (формализация задачи)	3
4 Спецификация программы	3
5 Формирование датасетов	3
6 Предварительный анализ данных	4
7 Внесение пропусков	4
8 Заполнение пропусков	4
9 Оценка результатов	4
10 Сравнение эффективности алгоритмов	4
11 Анализ результатов работы алгоритма	4
12 Контрольный пример	5
13 Блок-схема программы	7
14 Контрольный пример	7
15 Вывод	8
16 Источники	9

1 Цель работы

Изучить и реализовать методы заполнения пропусков в датасетах и определить их эффективность, сравнивая друг с другом.

2 Использованные методы импутации:

- Заполнение средним значением (mean)
- Заполнение медианой (median)
- Заполнение модой (mode)
- Заполнение предыдущими значениями (fill)
- Hot-deck
- Линейная регрессия (Linear reg.)
- Стохастические методы
- Сплайн-интерполяция

3 Описание задачи (формализация задачи)

Создать систему, которая:

- Восстанавливает лучшие значения из датасета
- Выбирает наилучший метод заполнения данных для данного датасета
- Выводит отчет о методах, оценках, метриках и времени выполнения

4 Спецификация программы

Входные данные: датасет в формате CSV. Можно выбрать все файлы из указанной пользователем папки.

Выходные данные:

- Наилучший метод
 - График всех методов
 - График ошибок для каждого метода в зависимости от процента пропуска
- Проект состоит из следующих основных модулей:
- `main.py` — основной файл для запуска анализа
 - `data_loading.py` — загрузка данных
 - `data_preprocessing.py` — предварительная обработка данных
 - `imputation_methods.py` — реализация методов импутации
 - `evaluation.py` — оценка методов
 - `visualization.py` — визуализация результатов

5 Формирование датасетов

Для исследования были сформированы три датасета:

- Малый (~ 1000–10 000 записей)
- Средний (~ 20 000–75 000 записей)
- Большой (~ 100 000–250 000 записей)

Данные генерировались с использованием программы из первой лабораторной работы прошлого семестра. Каждый датасет содержал числовые значения, характеризующие определённое распределение (например, нормальное, равномерное и т.д.).

6 Предварительный анализ данных

Для каждого датасета были вычислены следующие статистические характеристики:

- Среднее значение
- Медиана
- Мода
- Построены графики распределения (гистограммы)

Эти значения стали эталонными для дальнейшего сравнения.

7 Внесение пропусков

В каждом датасете искусственно удалялись значения в столбцах: 3%, 5%, 10%, 20%, 30% от общего числа записей. Удаление проводилось таким образом, чтобы создать выбросы — то есть удаление не было случайным, а затрагивало экстремальные значения.

8 Заполнение пропусков

Применялись следующие методы заполнения пропусков, описанные ранее.

9 Оценка результатов

После заполнения пропусков каждым из методов повторно вычислялись:

- Среднее значение
- Медиана
- Мода
- Графики распределения

Затем была рассчитана суммарная относительная погрешность для каждого метода.

10 Сравнение эффективности алгоритмов

Для оценки эффективности методов:

- Был сформирован эталонный датасет без пропусков
- На его основе были созданы датасеты с искусственными пропусками разной плотности (от 3% до 30%)
 - Для каждого уровня пропусков применялись перечисленные методы
 - Результаты сравнивались с эталоном по метрике суммарной относительной погрешности
 - Выявлен наиболее эффективный метод для каждой ситуации

11 Анализ результатов работы алгоритма

Результаты выполнения анализа на большом датасете (100 тыс. строк). По графику 1 и рисунку 2 можно видеть, что метод заполнения медианными значениями показывает наименьшую ошибку. Однако на реальных данных, где есть более реальная зависимость между признаками, стохастическая регрессия скорее всего показала бы наиболее правдоподобные результаты.

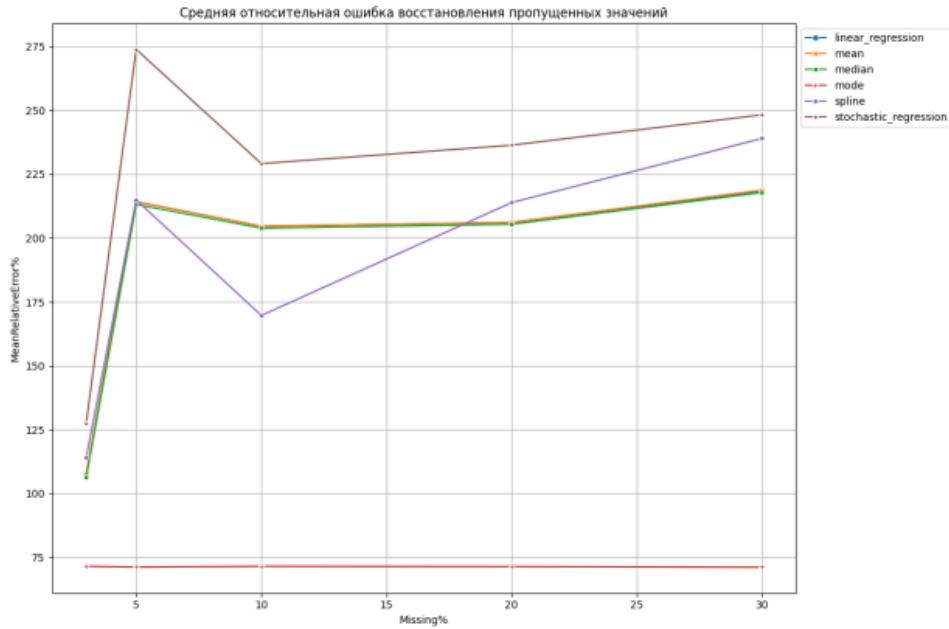


Рис. 1: Результаты анализа большого датасета

Method	Missing%	Column	MeanRelativeError%	NumEvaluated
mode	3	Unnamed: 0	62.454519241859636	9000.0
mode	5	Unnamed: 0	62.33910897011458	15000.0
mode	10	card_number	62.337729228800455	30000.0
mode	20	card_number	62.29517857272636	60000.0
mode	30	Unnamed: 0	62.284491790993044	90000.0

Рис. 2: Средняя относительная ошибка пропущенных значений

12 Контрольный пример

По итогу анализа выявлено, что заполнение пропусков модой является наиболее эффективным методом, если опираться на ошибку.

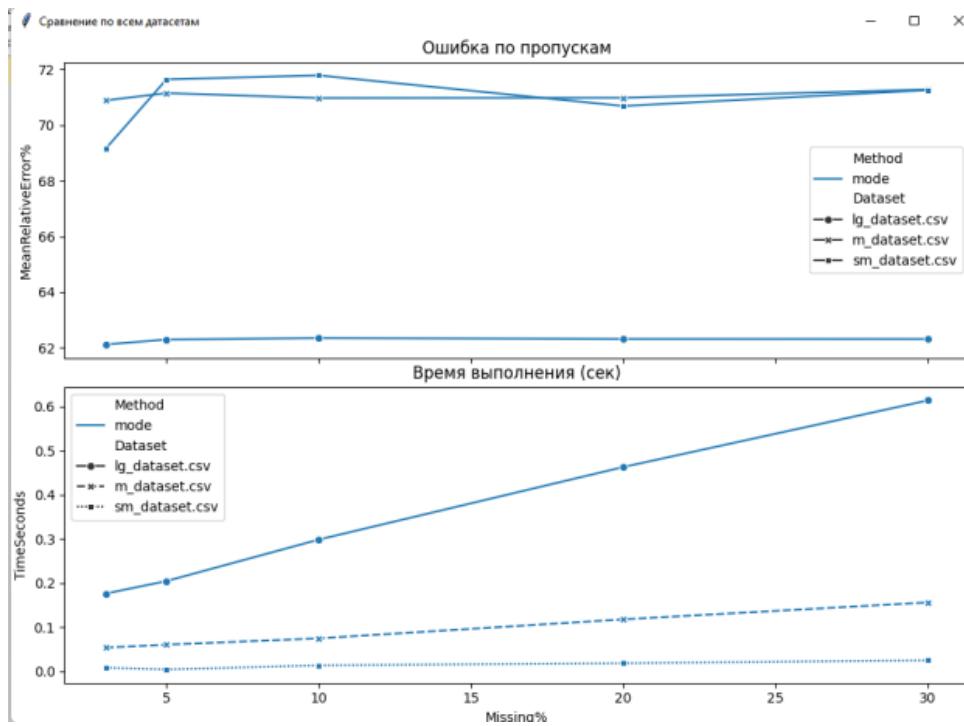


Рис. 3: График с лучшим методом по всем датасетам и затраченному времени

13 Блок-схема программы

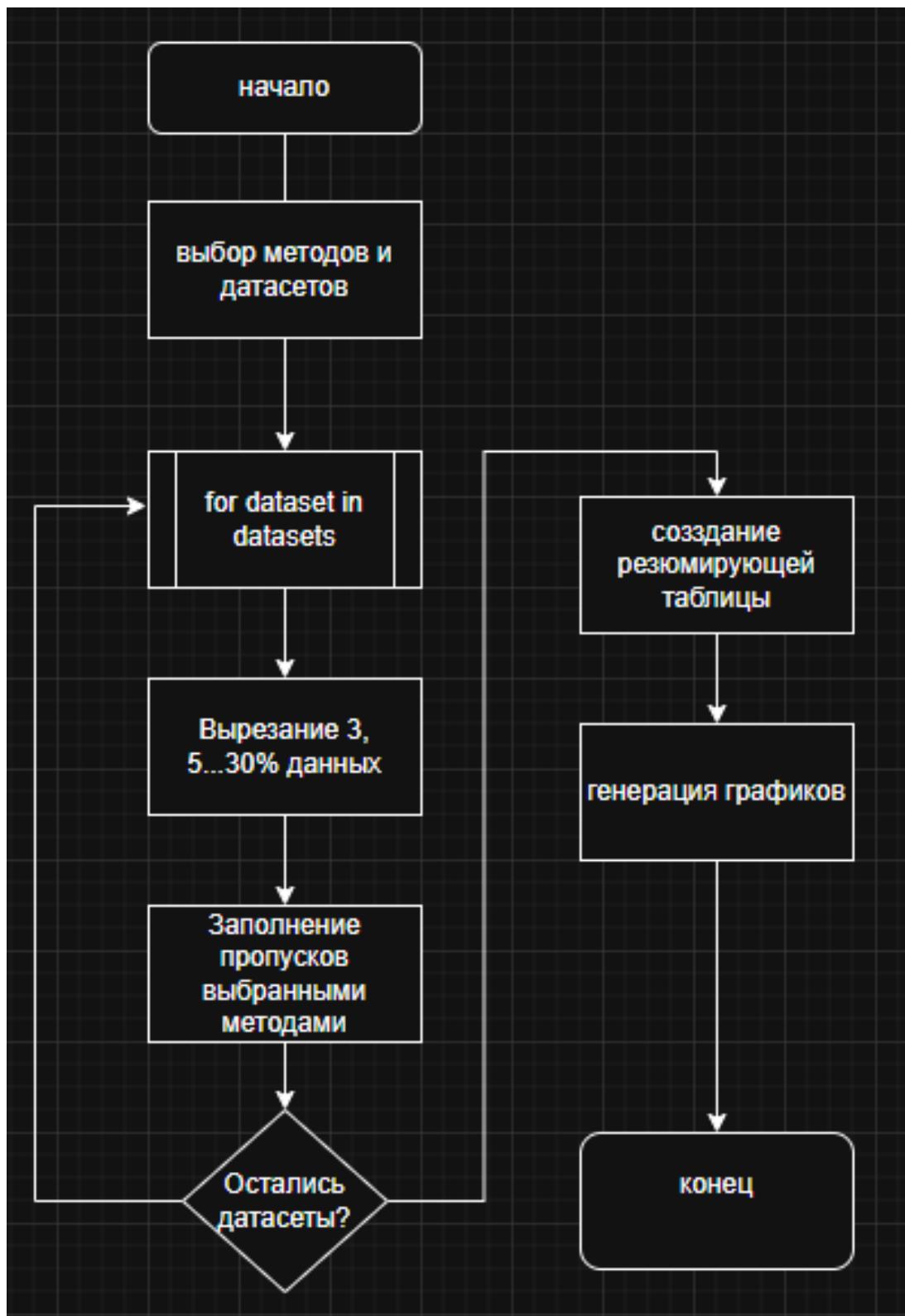


Рис. 4: Блок-схема основной программы

14 Контрольный пример

Интерфейс на начальном этапе позволяет:

- Загрузить CSV, выбрав файл из файловой системы
- Автоматически оценить папку, в которой лежат три датасета от маленького до большого по размеру
- Оценить конкретный датасет из этой папки, выбрав его номер из выпадающего списка (отсортированно от наименьшего к наибольшему)

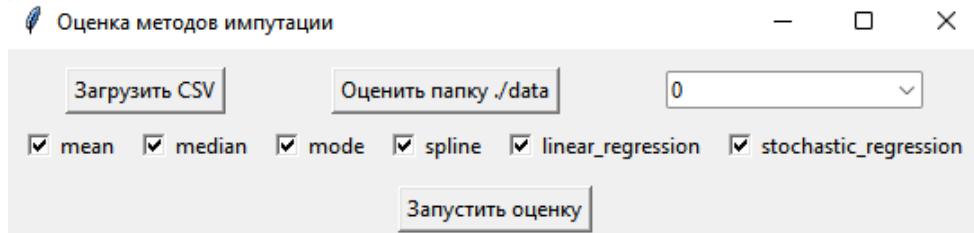


Рис. 5: Интерфейс

Пример выполнения анализа на маленьком датасете (15 тыс. строк до предобработки)

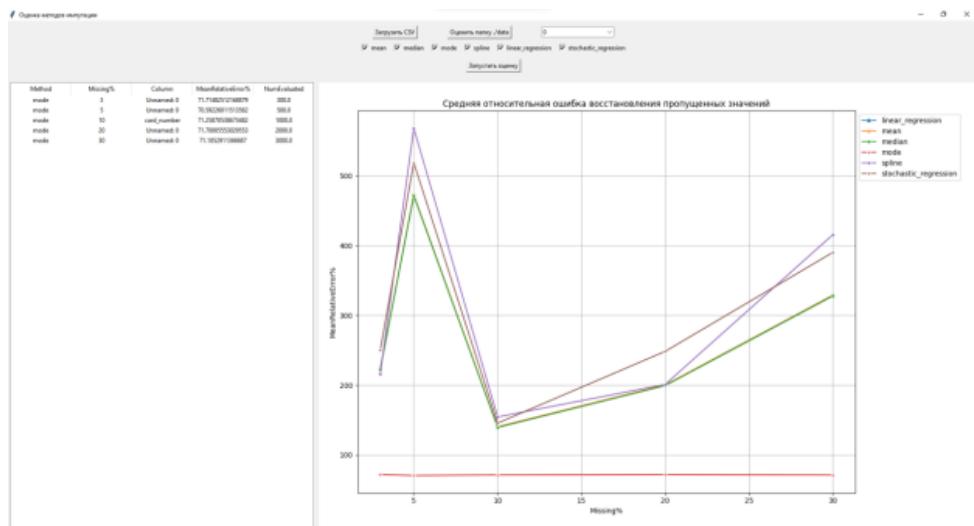


Рис. 6: Пример решения графа

15 Вывод

1. Все исследованные методы заполнения пропусков в большей или меньшей степени позволяют сохранить исходные статистические характеристики, однако их эффективность зависит от типа данных и объема выборки.
2. Наиболее универсальными и точными оказались методы, основанные на регрессии (особенно стохастическая), а также Хот-Дек и сплайн-интерполяция.
3. Простые методы (заполнение средним/медианой/модой) дают приемлемые результаты только при небольшом количестве пропусков.
4. Полученные результаты могут быть использованы для автоматического выбора метода обработки пропусков в зависимости от структуры и размера датасета.

16 Источники

1. Грабнер А. Н., Лялин В. С., Чилихин А. А. (2019). Обнаружение и коррекция аномалий и пропусков в данных. Научно-технический вестник информационных технологий, механики и оптики, № 6(124), с. 1024–1031. — Исследование методов работы с пропусками и выбросами в реальных наборах данных.
2. Python Software Foundation. (2023). Pandas Documentation. <https://pandas.pydata.org/pandas-docs/stable/>
3. R Core Team (2023). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>
4. IBM Knowledge Center. (2023). Missing value handling in SPSS. <https://www.ibm.com/docs/en/spss-statistics>
5. Tkinter библиотека: <https://tkinter.org/>
6. Math библиотека: <https://math.org/>