**САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ**

**Факультет прикладной математики-процессов управления**

**Программа бакалавриата**

**“Большие данные и распределенная цифровая платформа”**

**ОТЧЕТ**

**по лабораторной работе №9**

**по дисциплине «Алгоритмы и структуры данных»**

**на тему «Заполнение пропусков**»

**Студент гр. 23Б15-пу**

**Бек В.А.**

**Преподаватель**

**Дик А.Г.**

**Санкт-Петербург**

**2025 г.**

Оглавление

1. [Цель работы 3](#_Toc195371756)
2. [Описание задачи (формализация задачи) 4](#_Toc195371757)
3. [Теоретическая часть 5](#_Toc195371758)

[**Задача коммивояжера:** 5](#_Toc195371759)

[**Муравьиный алгоритм:** 5](#_Toc195371760)

[**Модификация (Блуждающая колония):** 6](#_Toc195371761)

1. [Основные шаги программы 7](#_Toc195371762)
2. [Блок схема программы 9](#_Toc195371763)
3. [Описание программы 11](#_Toc195371764)
4. [Рекомендации пользователя 13](#_Toc195371765)
5. [Рекомендации программиста 15](#_Toc195371766)
6. [Контрольный пример 16](#_Toc195371767)
7. [Анализ работы алгоритма с модификацией и без 18](#_Toc195371768)
8. [Сравнение с другими методами решения задачи коммивояжёра 19](#_Toc195371769)
9. [Вывод 21](#_Toc195371770)
10. [Источники 22](#_Toc195371771)

# Цель работы

**Цель работы** — изучение и реализация различных методов заполнения пропусков в датасетах при помощи различных методов, а также оценка их эффективности на основе сравнительного анализа. В ходе работы будут использованы такие методы, как: попарное удаление, метод заполнения моды, метод восстановления пропущенного значения на основе Zet-алгоритма.

# Описание задачи (формализация задачи)

В данной лабораторной работе требуется решить задачу восстановления пропущенных значений в датасетах с использованием различных методов. Задача формализована следующим образом:

1. **Создание датасетов**: Для начала формируются три датасета различных размеров (малый, средний и большой) при помощи первой лабораторной.
2. **Удаление значений**: В процессе работы необходимо удалить пропорционально 3%, 5%, 10%, 20% и 30% значений в случайных ячейках столбцов.
3. **Применение методов заполнения**: Для восстановления пропущенных значений применяются следующие методы:
   * **Попарное удаление** — В классическом методе необходимо выбрать два столбца и для них анализировать только не **nan** значения, но в данном случае все строки с пропущенными значениями полностью удаляются из датасета.
   * **Заполнение модой** — пропущенные значения заполняются наиболее часто встречающимся значением (модой) в соответствующем столбце.
   * **Zet-алгоритм** — данный метод использует Евклидово расстояние для нахождения строк, похожих на данную и заполнения пропусков для неё.
4. **Оценка качества заполнения**: Для каждого метода вычисляется суммарная погрешность предсказанных значений по сравнению с истинными значениями. Эффективность каждого метода будет оцениваться на основе данной погрешности для каждого столбца по отдельности и суммарной погрешности.

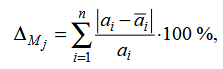
# Теоретическая часть

1. **Попарное удаление (Pairwise Deletion)** Этот метод заключается в удалении строк, в которых присутствуют пропущенные значения в любом из столбцов. Если говорить о классическом методе необходимо выбрать два столбца и для них анализировать только не **nan** значения
2. **Метод заполнения моды (Mode Imputation)** Метод заполнения пропусков с использованием наиболее часто встречающегося значения (моды) в столбце. Этот метод применим для категориальных данных, где отсутствующие значения заменяются наиболее распространенным значением среди остальных. Он помогает поддержать стабильность распределения данных, но не всегда точно отражает реальное распределение.
3. **На основе Zet-алгоритма:** Позволяет восстанавливать данные на основе сходства с уже имеющимися строками. Алгоритм основан на вычислении расстояний между строками в многомерном пространстве признаков, что позволяет более точно восстанавливать недостающие значения. В случае числовых данных пропуски заполняются на основе средней стоимости по соседним строкам, а для категориальных — на основе наиболее вероятных значений, вычисленных с учетом соседей.

### **Оценка качества восстановления данных:**

Для оценки качества восстановления данных используется метрика погрешности, которая рассчитывается как процент отклонения восстановленных значений от оригинальных:

**Погрешность для числовых данных:** Погрешность для числовых столбцов вычисляется как относительное отклонение между восстановленным и оригинальным значением. Для каждого числового столбца рассчитывается средняя погрешность по всем строкам, в которых данные были восстановлены. Формула расчета погрешности для каждого значения:



**Погрешность для категориальных данных**: Погрешность для категориальных данных измеряется как процент ошибок в восстановленных значениях, когда восстановленное значение не совпадает с оригинальным. Если в оригинальной строке было пропущено значение, то оно учитывается в расчете погрешности как 100%.

# Основные шаги программы

1. **Загрузка и отображение графического интерфейса**: С помощью библиотеки tkinter создаётся окно с интерфейсом, позволяющее пользователю выполнять основные действия.
2. **Выбор и загрузка файлов**: Интерфейс позволяет выбрать исходный файл данных, файл с пропусками и файл для восстановления, а также задать процент пропусков для удаления.
3. **Создание пропусков в данных**: Программа случайным образом удаляет заданный процент значений из Excel таблицы, в том числе блоками 4×4. Результат сохраняется в отдельный файл.
4. **Восстановление данных**: Пользователю предлагается выбор из нескольких методов восстановления:
   1. Попарное удаление
   2. Заполнение модой
   3. Zet-алгоритм

Восстановленные таблицы сохраняются в отдельные файлы.

1. **Расчёт погрешности восстановления**: Программа сравнивает восстановленные данные с оригинальными и выводит процентные ошибки по каждому столбцу, а также общее значение погрешности. Итоговая информация о погрешностях отображается в текстовом поле интерфейса.

# Блок схема программы

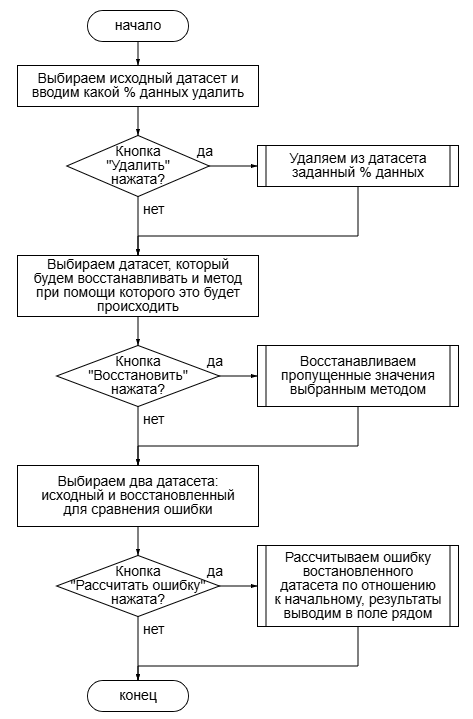
****

Рис. 1 Блок-схема основной программы

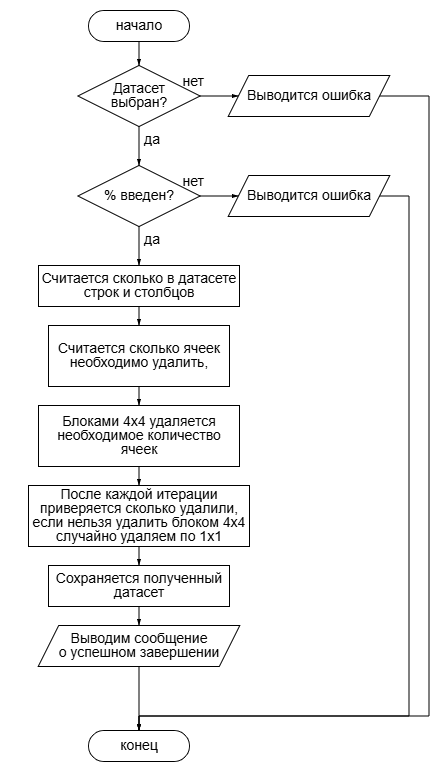


Рис. 2 Блок-схема подпрограммы удаления % значений

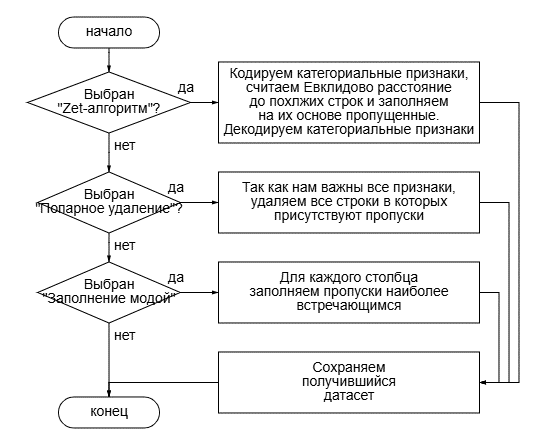


Рис. 3 Блок-схема подпрограммы восстановления пропущенных значений

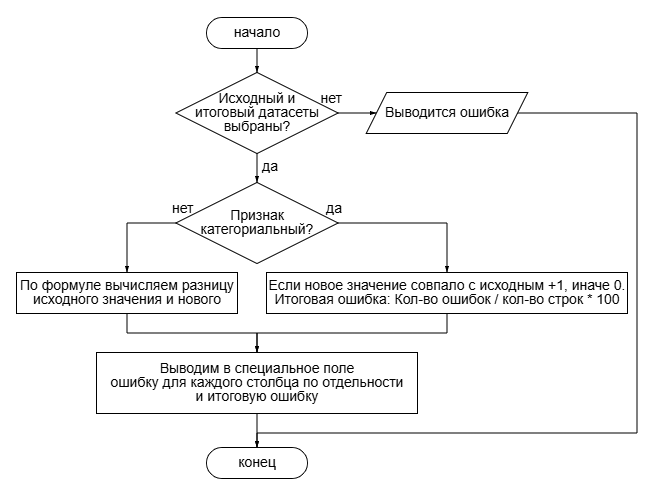


Рис. 4 Блок-схема подпрограммы подсчета ошибки восстановления

# Описание программы

Программная реализация задачи коммивояжёра написана на языке **Python 3.12.7** с использованием библиотек **tkinter, pandas, sklearn, numpy, pandas**. Программа представляет собой графический интерфейс, который позволяет удалять заданный % данных из датасета, восстанавливать пропущенные значения тремя способами, сравнивать результаты ошибки восстановления. Также предусмотрены отображение итогового результата ошибки для каждого столбца и всех сразу в специальном поле.

Таблица 1. restoration\_datachets.py

|  |  |  |
| --- | --- | --- |
| Функция | Описание | Возвращаемое значение |
| create\_gui() | Создаёт графический интерфейс. | None |
| select\_source\_file() | Открывает окно для выбора исходного файла и обновляет путь в поле ввода. | None |
| delete\_data() | Удаляет случайные значения из исходного файла в соответствии с введённым процентом. | None |
| select\_restore\_file() | Открывает окно для выбора файла для восстановления и обновляет путь в поле ввода. | None |
| recover\_data() | Восстанавливает данные датасета с использованием выбранного метода восстановления. | Восстановленный датасет |
| restore\_data() | Загружает данные из выбранного файла для восстановления, применяет метод восстановления и сохраняет результат в новый файл. | None |
| select\_input\_file\_name() | Открывает окно для выбора исходного файла для вычисления ошибки и обновляет путь в поле ввода. | None |
| select\_output\_file\_name() | Открывает окно для выбора восстановленного файла и обновляет путь в поле ввода. | None |

# Рекомендации пользователя

Для корректного выполнения программы следуйте приведённым шагам:

**Удаление данных (создание пропусков):**

1. **Нажмите кнопку «Выбрать»** напротив поля **«Исходный файл»** и выберите Excel-файл с данными.
2. Введите в поле **«Удалить (%)»** процент ячеек, которые нужно удалить.
3. Нажмите кнопку **«Удалить»** — программа создаст копию файла с пропущенными данными (\_nan.xlsx в конце имени файла).

**Восстановление данных:**

1. **Нажмите кнопку «Выбрать»** напротив поля **«Файл для восстановления»** и выберите файл с пропущенными значениями (например, созданный на предыдущем шаге).
2. В выпадающем списке **«Метод восстановления»** выберите один из трёх способов:
   * **Попарное удаление** — удалит строки с пропусками.
   * **Метод заполнения моды** — заполнит пропуски наиболее частыми значениями.
   * **На основе Zet-алгоритма** — восстановит значения на основе похожих строк.
3. Нажмите кнопку **«Восстановить»** — будет создан файл с восстановленными значениями (например, \_modefill.xlsx).

**Оценка качества восстановления:**

1. **Выберите файл ДО удаления данных** — нажмите **«Выбрать»** напротив **«Исходный файл»**.
2. **Выберите восстановленный файл** — нажмите **«Выбрать»** напротив **«Получившийся файл»**.
3. Нажмите кнопку **«Рассчитать ошибку»**.
4. В правой части окна появится подробная оценка ошибок восстановления по столбцам и в целом.

# Рекомендации программиста

Для корректного функционирования программы рекомендуется выполнить следующие действия:

1. **Установите необходимые библиотеки**:
   * Убедитесь, что у вас установлены библиотеки **tkinter, math, ttk** и **tkinter.simpledialog, sklearn**. **thinker** идет в стандартной поставке Python, но если она отсутствует, вы можете установить ее через пакетный менеджер вашей операционной системы.
2. **Проверьте версию Python**:
   * Рекомендуется использовать **Python** версии **3.1** или выше, чтобы избежать возможных проблем с совместимостью библиотек и функциональностью **tkinter**.
3. **Проверка функций и интерфейса**:
   * Убедитесь, что все элементы интерфейса (кнопки, текстовые поля и таблицы) работают корректно. Попробуйте вводить различные значения и проверить, что результаты рассчитываются и отображаются правильно.

**Код программы:**

**<https://github.com/Kliooo/Algorithms-and-data-structures>**

# Контрольный пример

* **Запуск программы:** Для запуска программы используйте файл restoration\_datachets.py. Программа откроет графический интерфейс (Рис. 5).

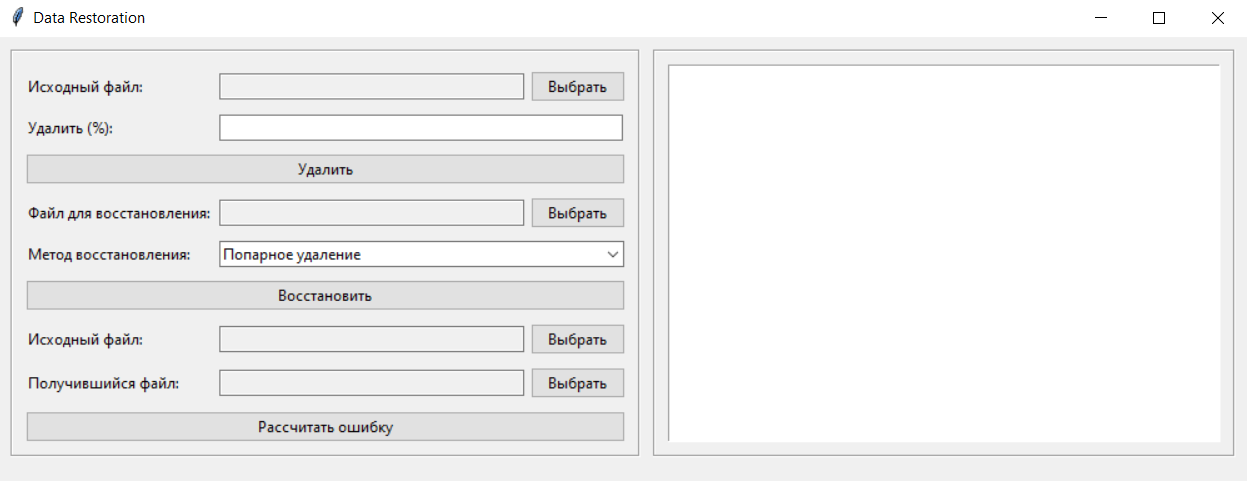


Рис. 5 Графический интерфейс программы

* **Удаление заданного % данных: Нажмите кнопку «Выбрать»** напротив поля **«Исходный файл»** и выберите Excel-файл с данными. ведите в поле **«Удалить (%)»** процент ячеек, которые нужно удалить. Нажмите кнопку **«Удалить»** — программа создаст копию файла с пропущенными данными (\_nan.xlsx в конце имени файла). (Рис. 6)

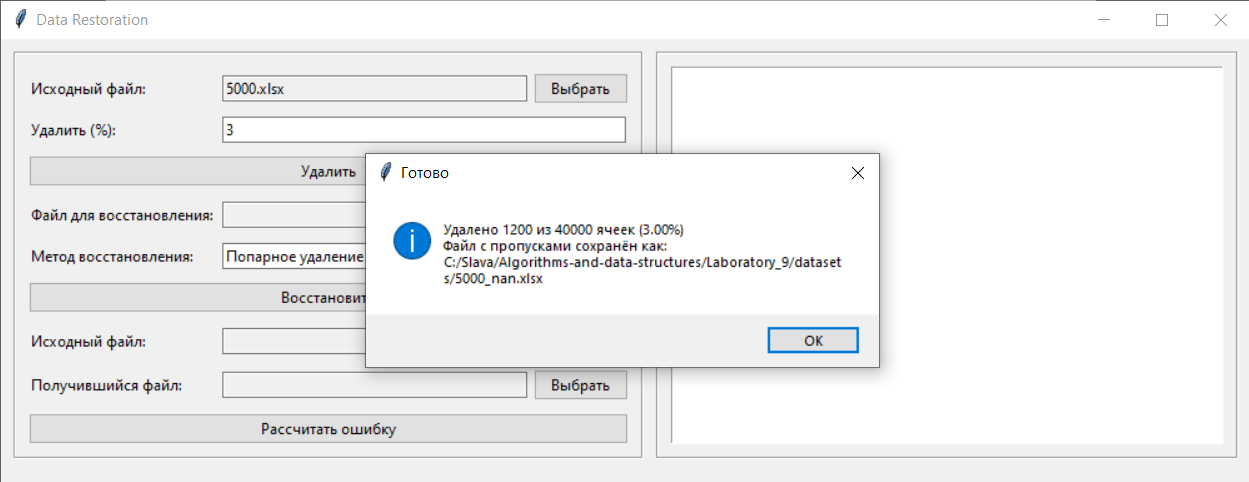


Рис. 6 Удаление части данных

* **Восстановление пропущенных значений:** **Нажмите кнопку «Выбрать»** напротив поля **«Файл для восстановления»** и выберите файл с пропущенными значениями (например, созданный на предыдущем шаге). Далее выберете один из трех методов и нажмите **«Восстановить». (Рис. 7)**

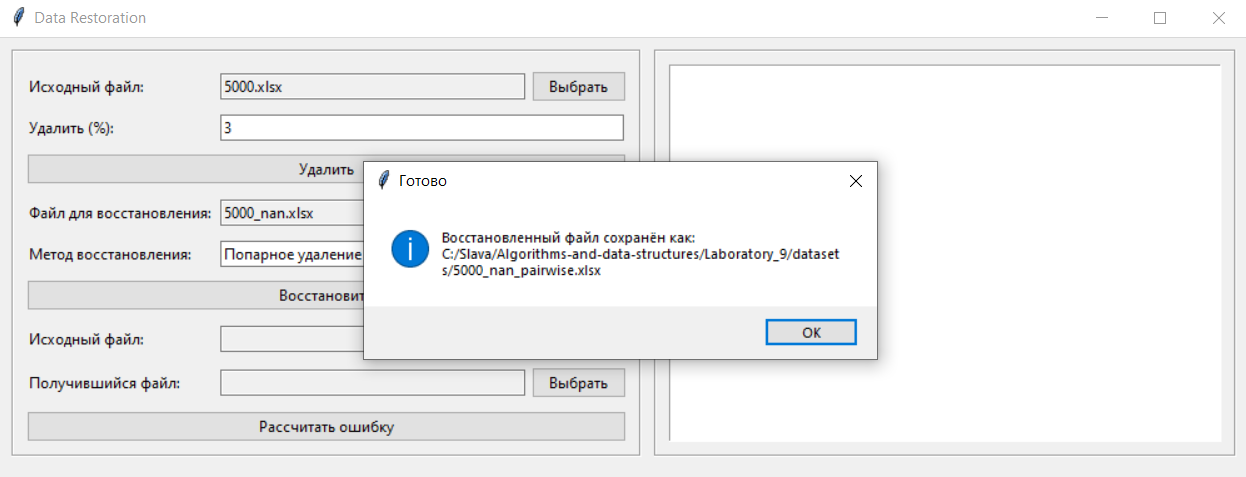


Рис. 7 Восстановление данных

* **Расчёт ошибки:** Выберите файл ДО удаления данных — нажмите «Выбрать» напротив «Исходный файл». Выберите восстановленный файл — нажмите «Выбрать» напротив «Получившийся файл». Нажмите кнопку «Рассчитать ошибку». В правой части окна (большое текстовое поле) появится подробная оценка ошибок восстановления по столбцам и в целом. (Рис. 8)

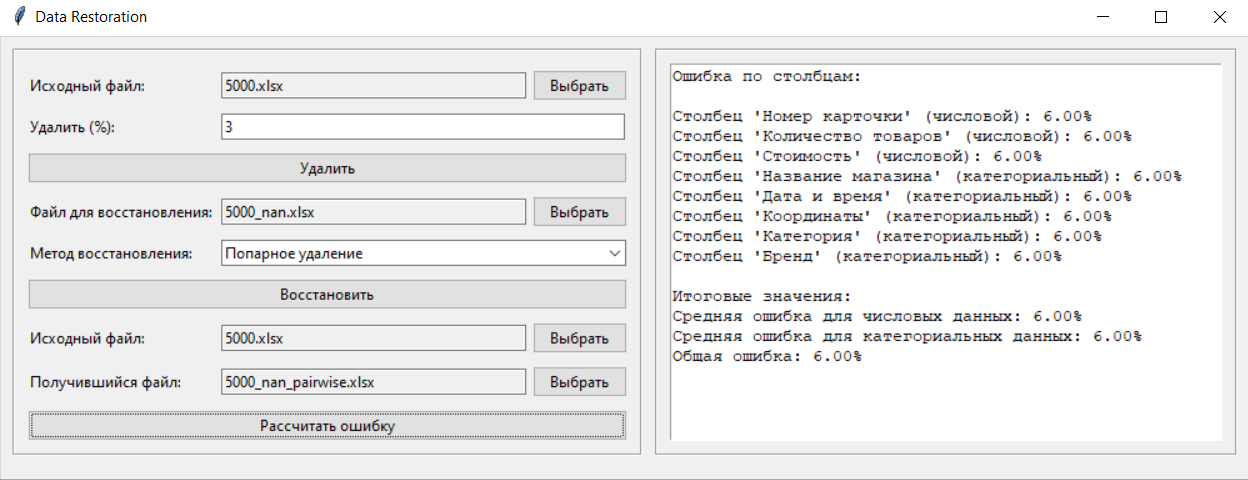


Рис. 8 Расчёт ошибки

# Анализ работы

Анализ методов восстановления датасета проводился на различных размерах датасетов и различных размерах датасетов и с разным %-ом удаляемых данных. Результаты приведены в таблице:

Таблица 2. Результаты работы

|  |  |  |  |
| --- | --- | --- | --- |
| Размер датасета | % удаления | Метод восстановления | Результаты |
| 5000 | 3 | Попарное удаление | Столбец 'Номер карточки' (числовой): 6.00%  Столбец 'Количество товаров' (числовой): 6.00%  Столбец 'Стоимость' (числовой): 6.00%  Столбец 'Название магазина' (категориальный): 6.00%  Столбец 'Дата и время' (категориальный): 6.00%  Столбец 'Координаты' (категориальный): 6.00%  Столбец 'Категория' (категориальный): 6.00%  Столбец 'Бренд' (категориальный): 6.00%  Средняя ошибка для числовых данных: 6.00%  Средняя ошибка для категориальных данных: 6.00%  Общая ошибка: 6.00% |
| 5000 | 3 | Заполнение модой | Столбец 'Номер карточки' (числовой): 1.34%  Столбец 'Количество товаров' (числовой): 0.43%  Столбец 'Стоимость' (числовой): 0.80%  Столбец 'Название магазина' (категориальный): 1.18%  Столбец 'Дата и время' (категориальный): 2.64%  Столбец 'Координаты' (категориальный): 3.96%  Столбец 'Категория' (категориальный): 4.84%  Столбец 'Бренд' (категориальный): 4.58%  Средняя ошибка для числовых данных: 0.86%  Средняя ошибка для категориальных данных: 3.44%  Общая ошибка: 2.15% |
| 5000 | 3 | Zet-алгоритм | Столбец 'Номер карточки' (числовой): 1.24%  Столбец 'Количество товаров' (числовой): 0.46%  Столбец 'Стоимость' (числовой): 35.36%  Столбец 'Название магазина' (категориальный): 1.26%  Столбец 'Дата и время' (категориальный): 2.64%  Столбец 'Координаты' (категориальный): 4.02%  Столбец 'Категория' (категориальный): 4.78%  Столбец 'Бренд' (категориальный): 4.64%  Средняя ошибка для числовых данных: 12.35%  Средняя ошибка для категориальных данных: 3.47%  Общая ошибка: 7.91% |
| 5000 | 5 | Попарное удаление | Столбец 'Номер карточки' (числовой): 9.98%  Столбец 'Количество товаров' (числовой): 9.98%  Столбец 'Стоимость' (числовой): 9.98%  Столбец 'Название магазина' (категориальный): 9.98%  Столбец 'Дата и время' (категориальный): 9.98%  Столбец 'Координаты' (категориальный): 9.98%  Столбец 'Категория' (категориальный): 9.98%  Столбец 'Бренд' (категориальный): 9.98%  Средняя ошибка для числовых данных: 9.98%  Средняя ошибка для категориальных данных: 9.98%  Общая ошибка: 9.98% |
| 5000 | 5 | Заполнение модой | Столбец 'Номер карточки' (числовой): 2.46%  Столбец 'Количество товаров' (числовой): 0.85%  Столбец 'Стоимость' (числовой): 1.68%  Столбец 'Название магазина' (категориальный): 1.88%  Столбец 'Дата и время' (категориальный): 4.08%  Столбец 'Координаты' (категориальный): 5.66%  Столбец 'Категория' (категориальный): 7.74%  Столбец 'Бренд' (категориальный): 7.70%  Средняя ошибка для числовых данных: 1.66%  Средняя ошибка для категориальных данных: 5.41%  Общая ошибка: 3.54% |
| 5000 | 5 | Zet-алгоритм | Столбец 'Номер карточки' (числовой): 2.08%  Столбец 'Количество товаров' (числовой): 0.83%  Столбец 'Стоимость' (числовой): 48.14%  Столбец 'Название магазина' (категориальный): 1.82%  Столбец 'Дата и время' (категориальный): 4.08%  Столбец 'Координаты' (категориальный): 5.82%  Столбец 'Категория' (категориальный): 7.86%  Столбец 'Бренд' (категориальный): 7.82%  Средняя ошибка для числовых данных: 17.02%  Средняя ошибка для категориальных данных: 5.48%  Общая ошибка: 11.25% |
| 5000 | 10 | Попарное удаление | Столбец 'Номер карточки' (числовой): 19.94%  Столбец 'Количество товаров' (числовой): 19.94%  Столбец 'Стоимость' (числовой): 19.94%  Столбец 'Название магазина' (категориальный): 19.94%  Столбец 'Дата и время' (категориальный): 19.94%  Столбец 'Координаты' (категориальный): 19.94%  Столбец 'Категория' (категориальный): 19.94%  Столбец 'Бренд' (категориальный): 19.94%  Средняя ошибка для числовых данных: 19.94%  Средняя ошибка для категориальных данных: 19.94%  Общая ошибка: 19.94% |
| 5000 | 10 | Заполнение модой | Столбец 'Номер карточки' (числовой): 5.15%  Столбец 'Количество товаров' (числовой): 1.83%  Столбец 'Стоимость' (числовой): 3.04%  Столбец 'Название магазина' (категориальный): 3.58%  Столбец 'Дата и время' (категориальный): 7.60%  Столбец 'Координаты' (категориальный): 10.90%  Столбец 'Категория' (категориальный): 15.48%  Столбец 'Бренд' (категориальный): 15.72%  Средняя ошибка для числовых данных: 3.34%  Средняя ошибка для категориальных данных: 10.66%  Общая ошибка: 7.00% |
| 5000 | 10 | Zet-алгоритм | Столбец 'Номер карточки' (числовой): 4.41%  Столбец 'Количество товаров' (числовой): 1.95%  Столбец 'Стоимость' (числовой): 137.23%  Столбец 'Название магазина' (категориальный): 3.58%  Столбец 'Дата и время' (категориальный): 7.60%  Столбец 'Координаты' (категориальный): 11.04%  Столбец 'Категория' (категориальный): 15.46%  Столбец 'Бренд' (категориальный): 15.82%  Средняя ошибка для числовых данных: 47.86%  Средняя ошибка для категориальных данных: 10.70%  Общая ошибка: 29.28% |
| 5000 | 20 | Попарное удаление | Столбец 'Номер карточки' (числовой): 39.00%  Столбец 'Количество товаров' (числовой): 39.00%  Столбец 'Стоимость' (числовой): 39.00%  Столбец 'Название магазина' (категориальный): 39.00%  Столбец 'Дата и время' (категориальный): 39.00%  Столбец 'Координаты' (категориальный): 39.00%  Столбец 'Категория' (категориальный): 39.00%  Столбец 'Бренд' (категориальный): 39.00%  Средняя ошибка для числовых данных: 39.00%  Средняя ошибка для категориальных данных: 39.00%  Общая ошибка: 39.00% |
| 5000 | 20 | Заполнение модой | Столбец 'Номер карточки' (числовой): 9.44%  Столбец 'Количество товаров' (числовой): 3.30%  Столбец 'Стоимость' (числовой): 6.58%  Столбец 'Название магазина' (категориальный): 8.46%  Столбец 'Дата и время' (категориальный): 17.12%  Столбец 'Координаты' (категориальный): 23.64%  Столбец 'Категория' (категориальный): 30.52%  Столбец 'Бренд' (категориальный): 30.12%  Средняя ошибка для числовых данных: 6.44%  Средняя ошибка для категориальных данных: 21.97%  Общая ошибка: 14.21% |
| 5000 | 20 | Zet-алгоритм | Столбец 'Номер карточки' (числовой): 8.13%  Столбец 'Количество товаров' (числовой): 3.21%  Столбец 'Стоимость' (числовой): 350.10%  Столбец 'Название магазина' (категориальный): 8.56%  Столбец 'Дата и время' (категориальный): 17.12%  Столбец 'Координаты' (категориальный): 24.24%  Столбец 'Категория' (категориальный): 30.56%  Столбец 'Бренд' (категориальный): 30.24%  Средняя ошибка для числовых данных: 120.48%  Средняя ошибка для категориальных данных: 22.14%  Общая ошибка: 71.31% |
| 5000 | 30 | Попарное удаление | Столбец 'Номер карточки' (числовой): 57.22%  Столбец 'Количество товаров' (числовой): 57.22%  Столбец 'Стоимость' (числовой): 57.22%  Столбец 'Название магазина' (категориальный): 57.22%  Столбец 'Дата и время' (категориальный): 57.22%  Столбец 'Координаты' (категориальный): 57.22%  Столбец 'Категория' (категориальный): 57.22%  Столбец 'Бренд' (категориальный): 57.22%  Средняя ошибка для числовых данных: 57.22%  Средняя ошибка для категориальных данных: 57.22%  Общая ошибка: 57.22% |
| 5000 | 30 | Заполнение модой | Столбец 'Номер карточки' (числовой): 14.31%  Столбец 'Количество товаров' (числовой): 5.25%  Столбец 'Стоимость' (числовой): 10.66%  Столбец 'Название магазина' (категориальный): 12.94%  Столбец 'Дата и время' (категориальный): 24.80%  Столбец 'Координаты' (категориальный): 34.74%  Столбец 'Категория' (категориальный): 45.30%  Столбец 'Бренд' (категориальный): 44.38%  Средняя ошибка для числовых данных: 10.07%  Средняя ошибка для категориальных данных: 32.43%  Общая ошибка: 21.25% |
| 5000 | 30 | Zet-алгоритм | Столбец 'Номер карточки' (числовой): 13.28%  Столбец 'Количество товаров' (числовой): 5.17%  Столбец 'Стоимость' (числовой): 448.51%  Столбец 'Название магазина' (категориальный): 13.38%  Столбец 'Дата и время' (категориальный): 24.80%  Столбец 'Координаты' (категориальный): 35.24%  Столбец 'Категория' (категориальный): 45.34%  Столбец 'Бренд' (категориальный): 44.78%  Средняя ошибка для числовых данных: 155.66%  Средняя ошибка для категориальных данных: 32.71%  Общая ошибка: 94.18% |
| 50000 | 3 | Попарное удаление | Столбец 'Номер карточки' (числовой): 5.99%  Столбец 'Количество товаров' (числовой): 5.99%  Столбец 'Стоимость' (числовой): 5.99%  Столбец 'Название магазина' (категориальный): 5.99%  Столбец 'Дата и время' (категориальный): 5.99%  Столбец 'Координаты' (категориальный): 5.99%  Столбец 'Категория' (категориальный): 5.99%  Столбец 'Бренд' (категориальный): 5.99%  Средняя ошибка для числовых данных: 5.99%  Средняя ошибка для категориальных данных: 5.99%  Общая ошибка: 5.99% |
| 50000 | 3 | Заполнение модой | Столбец 'Номер карточки' (числовой): 1.46%  Столбец 'Количество товаров' (числовой): 0.51%  Столбец 'Стоимость' (числовой): 0.96%  Столбец 'Название магазина' (категориальный): 1.10%  Столбец 'Дата и время' (категориальный): 2.35%  Столбец 'Координаты' (категориальный): 3.53%  Столбец 'Категория' (категориальный): 4.60%  Столбец 'Бренд' (категориальный): 4.68%  Средняя ошибка для числовых данных: 0.98%  Средняя ошибка для категориальных данных: 3.25%  Общая ошибка: 2.12% |
| 50000 | 3 | Zet-алгоритм | Столбец 'Номер карточки' (числовой): 1.33%  Столбец 'Количество товаров' (числовой): 0.48%  Столбец 'Стоимость' (числовой): 81.62%  Столбец 'Название магазина' (категориальный): 1.11%  Столбец 'Дата и время' (категориальный): 2.35%  Столбец 'Координаты' (категориальный): 3.52%  Столбец 'Категория' (категориальный): 4.60%  Столбец 'Бренд' (категориальный): 4.73%  Средняя ошибка для числовых данных: 27.81%  Средняя ошибка для категориальных данных: 3.26%  Общая ошибка: 15.53% |
| 50000 | 5 | Попарное удаление | Столбец 'Номер карточки' (числовой): 9.96%  Столбец 'Количество товаров' (числовой): 9.96%  Столбец 'Стоимость' (числовой): 9.96%  Столбец 'Название магазина' (категориальный): 9.96%  Столбец 'Дата и время' (категориальный): 9.96%  Столбец 'Координаты' (категориальный): 9.96%  Столбец 'Категория' (категориальный): 9.96%  Столбец 'Бренд' (категориальный): 9.96%  Средняя ошибка для числовых данных: 9.96%  Средняя ошибка для категориальных данных: 9.96%  Общая ошибка: 9.96% |
| 50000 | 5 | Заполнение модой | Столбец 'Номер карточки' (числовой): 2.43%  Столбец 'Количество товаров' (числовой): 0.96%  Столбец 'Стоимость' (числовой): 1.77%  Столбец 'Название магазина' (категориальный): 1.72%  Столбец 'Дата и время' (категориальный): 3.94%  Столбец 'Координаты' (категориальный): 5.49%  Столбец 'Категория' (категориальный): 7.44%  Столбец 'Бренд' (категориальный): 7.86%  Средняя ошибка для числовых данных: 1.72%  Средняя ошибка для категориальных данных: 5.29%  Общая ошибка: 3.51% |
| 50000 | 5 | Zet-алгоритм | Столбец 'Номер карточки' (числовой): 2.19%  Столбец 'Количество товаров' (числовой): 0.94%  Столбец 'Стоимость' (числовой): 186.29%  Столбец 'Название магазина' (категориальный): 1.76%  Столбец 'Дата и время' (категориальный): 3.94%  Столбец 'Координаты' (категориальный): 5.52%  Столбец 'Категория' (категориальный): 7.42%  Столбец 'Бренд' (категориальный): 7.99%  Средняя ошибка для числовых данных: 63.14%  Средняя ошибка для категориальных данных: 5.32%  Общая ошибка: 34.23% |
| 50000 | 10 | Попарное удаление | Столбец 'Номер карточки' (числовой): 19.84%  Столбец 'Количество товаров' (числовой): 19.84%  Столбец 'Стоимость' (числовой): 19.84%  Столбец 'Название магазина' (категориальный): 19.84%  Столбец 'Дата и время' (категориальный): 19.84%  Столбец 'Координаты' (категориальный): 19.84%  Столбец 'Категория' (категориальный): 19.84%  Столбец 'Бренд' (категориальный): 19.84%  Средняя ошибка для числовых данных: 19.84%  Средняя ошибка для категориальных данных: 19.84%  Общая ошибка: 19.84% |
| 50000 | 10 | Заполнение модой | Столбец 'Номер карточки' (числовой): 4.74%  Столбец 'Количество товаров' (числовой): 1.62%  Столбец 'Стоимость' (числовой): 3.14%  Столбец 'Название магазина' (категориальный): 3.92%  Столбец 'Дата и время' (категориальный): 8.37%  Столбец 'Координаты' (категориальный): 12.21%  Столбец 'Категория' (категориальный): 15.43%  Столбец 'Бренд' (категориальный): 15.25%  Средняя ошибка для числовых данных: 3.17%  Средняя ошибка для категориальных данных: 11.04%  Общая ошибка: 7.10% |
| 50000 | 10 | Zet-алгоритм | Столбец 'Номер карточки' (числовой): 4.18%  Столбец 'Количество товаров' (числовой): 1.63%  Столбец 'Стоимость' (числовой): 207.79%  Столбец 'Название магазина' (категориальный): 3.88%  Столбец 'Дата и время' (категориальный): 8.37%  Столбец 'Координаты' (категориальный): 12.23%  Столбец 'Категория' (категориальный): 15.42%  Столбец 'Бренд' (категориальный): 15.45%  Средняя ошибка для числовых данных: 71.20%  Средняя ошибка для категориальных данных: 11.07%  Общая ошибка: 41.14% |
| 50000 | 20 | Попарное удаление | Столбец 'Номер карточки' (числовой): 39.08%  Столбец 'Количество товаров' (числовой): 39.08%  Столбец 'Стоимость' (числовой): 39.08%  Столбец 'Название магазина' (категориальный): 39.08%  Столбец 'Дата и время' (категориальный): 39.08%  Столбец 'Координаты' (категориальный): 39.08%  Столбец 'Категория' (категориальный): 39.08%  Столбец 'Бренд' (категориальный): 39.08%  Средняя ошибка для числовых данных: 39.08%  Средняя ошибка для категориальных данных: 39.08%  Общая ошибка: 39.08% |
| 50000 | 20 | Заполнение модой | Столбец 'Номер карточки' (числовой): 9.66%  Столбец 'Количество товаров' (числовой): 3.54%  Столбец 'Стоимость' (числовой): 6.78%  Столбец 'Название магазина' (категориальный): 7.31%  Столбец 'Дата и время' (категориальный): 16.16%  Столбец 'Координаты' (категориальный): 23.13%  Столбец 'Категория' (категориальный): 30.13%  Столбец 'Бренд' (категориальный): 31.15%  Средняя ошибка для числовых данных: 6.66%  Средняя ошибка для категориальных данных: 21.58%  Общая ошибка: 14.12% |
| 50000 | 20 | Zet-алгоритм | Столбец 'Номер карточки' (числовой): 8.59%  Столбец 'Количество товаров' (числовой): 3.37%  Столбец 'Стоимость' (числовой): 583.63%  Столбец 'Название магазина' (категориальный): 7.33%  Столбец 'Дата и время' (категориальный): 16.16%  Столбец 'Координаты' (категориальный): 23.10%  Столбец 'Категория' (категориальный): 30.17%  Столбец 'Бренд' (категориальный): 31.42%  Средняя ошибка для числовых данных: 198.53%  Средняя ошибка для категориальных данных: 21.64%  Общая ошибка: 110.08% |
| 50000 | 30 | Попарное удаление | Столбец 'Номер карточки' (числовой): 57.25%  Столбец 'Количество товаров' (числовой): 57.25%  Столбец 'Стоимость' (числовой): 57.25%  Столбец 'Название магазина' (категориальный): 57.25%  Столбец 'Дата и время' (категориальный): 57.25%  Столбец 'Координаты' (категориальный): 57.25%  Столбец 'Категория' (категориальный): 57.25%  Столбец 'Бренд' (категориальный): 57.25%  Средняя ошибка для числовых данных: 57.25%  Средняя ошибка для категориальных данных: 57.25%  Общая ошибка: 57.25% |
| 50000 | 30 | Заполнение модой | Столбец 'Номер карточки' (числовой): 14.64%  Столбец 'Количество товаров' (числовой): 5.76%  Столбец 'Стоимость' (числовой): 10.84%  Столбец 'Название магазина' (категориальный): 11.77%  Столбец 'Дата и время' (категориальный): 23.94%  Столбец 'Координаты' (категориальный): 34.01%  Столбец 'Категория' (категориальный): 44.53%  Столбец 'Бренд' (категориальный): 45.75%  Средняя ошибка для числовых данных: 10.41%  Средняя ошибка для категориальных данных: 32.00%  Общая ошибка: 21.21% |
| 50000 | 30 | Zet-алгоритм | Столбец 'Номер карточки' (числовой): 13.34%  Столбец 'Количество товаров' (числовой): 5.93%  Столбец 'Стоимость' (числовой): 1014.46%  Столбец 'Название магазина' (категориальный): 11.93%  Столбец 'Дата и время' (категориальный): 23.94%  Столбец 'Координаты' (категориальный): 33.92%  Столбец 'Категория' (категориальный): 44.62%  Столбец 'Бренд' (категориальный): 46.18%  Средняя ошибка для числовых данных: 344.58%  Средняя ошибка для категориальных данных: 32.12%  Общая ошибка: 188.35% |
| 120000 | 3 | Попарное удаление | Столбец 'Номер карточки' (числовой): 5.99%  Столбец 'Количество товаров' (числовой): 5.99%  Столбец 'Стоимость' (числовой): 5.99%  Столбец 'Название магазина' (категориальный): 5.99%  Столбец 'Дата и время' (категориальный): 5.99%  Столбец 'Координаты' (категориальный): 5.99%  Столбец 'Категория' (категориальный): 5.99%  Столбец 'Бренд' (категориальный): 5.99%  Средняя ошибка для числовых данных: 5.99%  Средняя ошибка для категориальных данных: 5.99%  Общая ошибка: 5.99% |
| 120000 | 3 | Заполнение модой | Столбец 'Номер карточки' (числовой): 1.43%  Столбец 'Количество товаров' (числовой): 0.70%  Столбец 'Стоимость' (числовой): 0.95%  Столбец 'Название магазина' (категориальный): 1.10%  Столбец 'Дата и время' (категориальный): 2.46%  Столбец 'Координаты' (категориальный): 3.48%  Столбец 'Категория' (категориальный): 4.61%  Столбец 'Бренд' (категориальный): 4.68%  Средняя ошибка для числовых данных: 1.03%  Средняя ошибка для категориальных данных: 3.27%  Общая ошибка: 2.15% |
| 120000 | 3 | Zet-алгоритм | Столбец 'Номер карточки' (числовой): 1.28%  Столбец 'Количество товаров' (числовой): 0.49%  Столбец 'Стоимость' (числовой): 109.64%  Столбец 'Название магазина' (категориальный): 1.08%  Столбец 'Дата и время' (категориальный): 2.46%  Столбец 'Координаты' (категориальный): 3.46%  Столбец 'Категория' (категориальный): 4.63%  Столбец 'Бренд' (категориальный): 4.74%  Средняя ошибка для числовых данных: 37.14%  Средняя ошибка для категориальных данных: 3.27%  Общая ошибка: 20.21% |
| 120000 | 5 | Попарное удаление | Столбец 'Номер карточки' (числовой): 9.96%  Столбец 'Количество товаров' (числовой): 9.96%  Столбец 'Стоимость' (числовой): 9.96%  Столбец 'Название магазина' (категориальный): 9.96%  Столбец 'Дата и время' (категориальный): 9.96%  Столбец 'Координаты' (категориальный): 9.96%  Столбец 'Категория' (категориальный): 9.96%  Столбец 'Бренд' (категориальный): 9.96%  Средняя ошибка для числовых данных: 9.96%  Средняя ошибка для категориальных данных: 9.96%  Общая ошибка: 9.96% |
| 120000 | 5 | Заполнение модой | Столбец 'Номер карточки' (числовой): 2.47%  Столбец 'Количество товаров' (числовой): 0.87%  Столбец 'Стоимость' (числовой): 1.55%  Столбец 'Название магазина' (категориальный): 1.80%  Столбец 'Дата и время' (категориальный): 3.97%  Столбец 'Координаты' (категориальный): 5.82%  Столбец 'Категория' (категориальный): 7.74%  Столбец 'Бренд' (категориальный): 7.85%  Средняя ошибка для числовых данных: 1.63%  Средняя ошибка для категориальных данных: 5.44%  Общая ошибка: 3.54% |
| 120000 | 5 | Zet-алгоритм | Столбец 'Номер карточки' (числовой): 2.15%  Столбец 'Количество товаров' (числовой): 0.82%  Столбец 'Стоимость' (числовой): 173.35%  Столбец 'Название магазина' (категориальный): 1.78%  Столбец 'Дата и время' (категориальный): 3.97%  Столбец 'Координаты' (категориальный): 5.82%  Столбец 'Категория' (категориальный): 7.74%  Столбец 'Бренд' (категориальный): 7.92%  Средняя ошибка для числовых данных: 58.77%  Средняя ошибка для категориальных данных: 5.44%  Общая ошибка: 32.11% |
| 120000 | 10 | Попарное удаление | Столбец 'Номер карточки' (числовой): 19.85%  Столбец 'Количество товаров' (числовой): 19.85%  Столбец 'Стоимость' (числовой): 19.85%  Столбец 'Название магазина' (категориальный): 19.85%  Столбец 'Дата и время' (категориальный): 19.85%  Столбец 'Координаты' (категориальный): 19.85%  Столбец 'Категория' (категориальный): 19.85%  Столбец 'Бренд' (категориальный): 19.85%  Средняя ошибка для числовых данных: 19.85%  Средняя ошибка для категориальных данных: 19.85%  Общая ошибка: 19.85% |
| 120000 | 10 | Заполнение модой | Столбец 'Номер карточки' (числовой): 4.99%  Столбец 'Количество товаров' (числовой): 1.69%  Столбец 'Стоимость' (числовой): 3.48%  Столбец 'Название магазина' (категориальный): 3.73%  Столбец 'Дата и время' (категориальный): 7.85%  Столбец 'Координаты' (категориальный): 11.26%  Столбец 'Категория' (категориальный): 15.01%  Столбец 'Бренд' (категориальный): 15.53%  Средняя ошибка для числовых данных: 3.38%  Средняя ошибка для категориальных данных: 10.67%  Общая ошибка: 7.03% |
| 120000 | 10 | Zet-алгоритм | Столбец 'Номер карточки' (числовой): 4.31%  Столбец 'Количество товаров' (числовой): 1.73%  Столбец 'Стоимость' (числовой): 407.09%  Столбец 'Название магазина' (категориальный): 3.75%  Столбец 'Дата и время' (категориальный): 7.85%  Столбец 'Координаты' (категориальный): 11.25%  Столбец 'Категория' (категориальный): 14.99%  Столбец 'Бренд' (категориальный): 15.65%  Средняя ошибка для числовых данных: 137.71%  Средняя ошибка для категориальных данных: 10.70%  Общая ошибка: 74.21% |
| 120000 | 20 | Попарное удаление | Столбец 'Номер карточки' (числовой): 39.15%  Столбец 'Количество товаров' (числовой): 39.15%  Столбец 'Стоимость' (числовой): 39.15%  Столбец 'Название магазина' (категориальный): 39.15%  Столбец 'Дата и время' (категориальный): 39.15%  Столбец 'Координаты' (категориальный): 39.15%  Столбец 'Категория' (категориальный): 39.15%  Столбец 'Бренд' (категориальный): 39.15%  Средняя ошибка для числовых данных: 39.15%  Средняя ошибка для категориальных данных: 39.15%  Общая ошибка: 39.15% |
| 120000 | 20 | Заполнение модой | Столбец 'Номер карточки' (числовой): 10.96%  Столбец 'Количество товаров' (числовой): 5.58%  Столбец 'Стоимость' (числовой): 5.65%  Столбец 'Название магазина' (категориальный): 8.81%  Столбец 'Дата и время' (категориальный): 15.32%  Столбец 'Координаты' (категориальный): 25.13%  Столбец 'Категория' (категориальный): 29.53%  Столбец 'Бренд' (категориальный): 34.62%  Средняя ошибка для числовых данных: 8.72%  Средняя ошибка для категориальных данных: 24.88%  Общая ошибка: 16.80% |
| 120000 | 20 | Zet-алгоритм | Столбец 'Номер карточки' (числовой): 9.43%  Столбец 'Количество товаров' (числовой): 2.98%  Столбец 'Стоимость' (числовой): 874.55%  Столбец 'Название магазина' (категориальный): 9.22%  Столбец 'Дата и время' (категориальный): 14.11%  Столбец 'Координаты' (категориальный): 25.17%  Столбец 'Категория' (категориальный): 33.11%  Столбец 'Бренд' (категориальный): 29.82%  Средняя ошибка для числовых данных: 250.19%  Средняя ошибка для категориальных данных: 24.24%  Общая ошибка: 137.22% |
| 120000 | 30 | Попарное удаление | Столбец 'Номер карточки' (числовой): 57.87%  Столбец 'Количество товаров' (числовой): 57.87%  Столбец 'Стоимость' (числовой): 57.87%  Столбец 'Название магазина' (категориальный): 57.87%  Столбец 'Дата и время' (категориальный): 57.87%  Столбец 'Координаты' (категориальный): 57.87%  Столбец 'Категория' (категориальный): 57.87%  Столбец 'Бренд' (категориальный): 57.87%  Средняя ошибка для числовых данных: 57.87%  Средняя ошибка для категориальных данных: 57.87%  Общая ошибка: 57.87% |
| 120000 | 30 | Заполнение модой | Столбец 'Номер карточки' (числовой): 16.34%  Столбец 'Количество товаров' (числовой): 5.23%  Столбец 'Стоимость' (числовой): 11.96%  Столбец 'Название магазина' (категориальный): 13.41%  Столбец 'Дата и время' (категориальный): 21.60%  Столбец 'Координаты' (категориальный): 30.21%  Столбец 'Категория' (категориальный): 46.11%  Столбец 'Бренд' (категориальный): 44.70%  Средняя ошибка для числовых данных: 13.22%  Средняя ошибка для категориальных данных: 34.19%  Общая ошибка: 23.71% |
| 120000 | 30 | Zet-алгоритм | Столбец 'Номер карточки' (числовой): 15.76%  Столбец 'Количество товаров' (числовой): 6.00%  Столбец 'Стоимость' (числовой): 1545.11%  Столбец 'Название магазина' (категориальный): 12.85%  Столбец 'Дата и время' (категориальный): 21.88%  Столбец 'Координаты' (категориальный): 35.75%  Столбец 'Категория' (категориальный): 48.42%  Столбец 'Бренд' (категориальный): 43.21%  Средняя ошибка для числовых данных: 522.29%  Средняя ошибка для категориальных данных: 32.42%  Общая ошибка: 228.62% |

**Первичные наблюдения:**

* **Попарное удаление** показывает **стабильную, но высокую ошибку**, примерно равную удвоенному % удаленных данных. Это достаточно ожидаемо, так как метод просто удаляет строки с пропусками, что не восстанавливает данные, а теряет их.
* **Заполнение модой** демонстрирует **наименьшую общую ошибку** на большинстве тестов, особенно при малом проценте пропусков.
* **Zet-алгоритм** показывает **низкую ошибку для категориальных данных и количества товаров,** но **сильно проваливается на числовых**, особенно в столбце **Стоимость**, где ошибка в ряде случаев достигает **сотен или даже тысяч процентов**, это, скорее всего, связано со способом оценки и разбросанностью данных в данном столбце (от 350 до 100000).

**Подробное сравнение по размеру датасета и проценту пропусков:**

**5000 строк:**

* 3 – 5% пропусков:
  + Попарное удаление — Общая ошибка = 2 \* процент пропусков
  + Мода — лучшая (общая ошибка: 2–3%)
  + Zet-алгоритм — приемлем, но нестабилен по числовым столбцам (до 48%), в остальном показывает себя примерно, как Мода.
* 10% пропусков:
  + Попарное удаление — Общая ошибка = 2 \* процент пропусков
  + Мода — лучшая (общая ошибка: 7%)
  + Zet-алгоритм — ошибка очевидно увеличивается до 11%
* 20 – 30% пропусков:
  + Попарное удаление — Общая ошибка = 2 \* процент пропусков
  + Мода продолжает оставаться лучшей, хоть и теряет точность (до +-20%)
  + Zet-алгоритм резко ухудшается (ошибка в числовых до +-**80%**)

**50000 строк:**

* 3 – 5% пропусков:
  + Попарное удаление — Общая ошибка = 2 \* процент пропусков
  + Мода — снова показывает хороший результат 2-4%
  + Zet-алгоритм — общая ошибка по столбцу Стоимость начинает расти и уже сейчас достигает 80% при 3% пропусков и 180% при 5% пропусков.
* 10% пропусков:
  + Попарное удаление — Общая ошибка = 2 \* процент пропусков
  + Мода — общая ошибка 7%
  + Zet-алгоритм — ошибка по столбцу Стоимость вырастает до 207%
* 20 – 30% пропусков:
  + Попарное удаление — Общая ошибка = 2 \* процент пропусков
  + Мода — общая ошибка от 14% до 21%
  + Zet-алгоритм — ошибка по столбцу Стоимость резко вырастает до 580% при 20% пропусков и 1000% при 30% пропусков

**120000 строк:**

* 3 – 5% пропусков:
  + Попарное удаление — Общая ошибка = 2 \* процент пропусков
  + Мода — показывает примерно такой же результат, как и при других количествах строк: 2-4%
  + Zet-алгоритм — ошибка по столбцу Стоимость остается такой же высокой: 100-200%
* 10% пропусков:
  + Попарное удаление — Общая ошибка = 2 \* процент пропусков
  + Мода — общая ошибка так и остается 7%
  + Zet-алгоритм — ошибка по столбцу Стоимость снова резко вырастает до 400%
* 20 – 30% пропусков:
  + Попарное удаление — Общая ошибка = 2 \* процент пропусков
  + Мода — общая ошибка от 16% до 23%
  + Zet-алгоритм — и снова ошибка по столбцу Стоимость резко вырастает до 870% при 20% пропущенных и 1500% при 30% пропущенных

**Вывод:**

* **Лучший метод**: Общий лучший результат показало заполнение модой, всегда выдавал приемлемую ошибку и не показывал серьезных выбросов.
* **Худший метод по устойчивости**:Zet-алгоритм, несмотря на очень интересный и теоретически эффективный способ заполнения пропусков дает самый худший результат из-за столбца: Стоимость. В остальном, если не рассматривать данный выброс, результат получается примерно, как у заполнения модой, но времени на заполнение тратится гораздо больше из-за сложности метода. Например, для 30% пропусков при 120000 строк понадобилось примерно 7-8 часов.
* **Попарное удаление** — самый простой из данных методов и не очень подходящий для задачи восстановления данных, всегда давал ошибку примерно равную удвоенному количеству удаленных данных.

# Вывод

В рамках данной работы были изучены и реализованы три метода заполнения пропусков в датасетах: попарное удаление, заполнение модой и Zet-алгоритм. Проведено сравнительное тестирование этих методов на датасетах различного размера 5000, 50000 и 120000 строк с разным процентом удалённых данных 3%, 5%, 10%, 20%, 30%.

Основные результаты:

* **Заполнение модой** продемонстрировало наилучшую общую точность, обеспечивая минимальную погрешность (2–23% в зависимости от процента пропусков и размера датасета). Метод оказался устойчивым и эффективным как для числовых, так и для категориальных данных, не показывая значительных выбросов.
* **Zet-алгоритм**, несмотря на теоретическую перспективность, показал нестабильные результаты из-за высокой погрешности в столбце «Стоимость» (до 1500% при 30% пропусков). Для категориальных данных и других числовых столбцов его точность была примерно сопоставима с заполнением модой, однако высокая вычислительная сложность делает его менее практичным.
* **Попарное удаление** оказалось наименее подходящим для восстановления данных, так как приводит к потере информации, а погрешность стабильно равна удвоенному проценту удалённых данных (6–57%).

Таким образом, заполнение модой является предпочтительным методом для большинства случаев благодаря своей простоте, скорости и стабильной точности. Zet-алгоритм требует доработки, особенно в обработке числовых данных с большим разбросом. Попарное удаление может быть использовано только в случаях, когда допустима потеря данных.

# Источники

* Редактор блок-схем.

[*https://programforyou.ru/block-diagram-redactor*](https://programforyou.ru/block-diagram-redactor)

*дата обращения: (28.04.2025)*

* tkinter — Библиотека для создания графических интерфейсов в Python.

[*https://docs.python.org/3/library/tkinter.html*](https://docs.python.org/3/library/tkinter.html%20)

*дата обращения: (28.04.2025)*

* math — Библиотека для работы с математическими функциями.

[*https://docs.python.org/3/library/math.html*](https://docs.python.org/3/library/math.html)

*дата обращения: (28.04.2025)*

* openpyxl — Библиотека для записи и чтения файлов Excel.

[*https://openpyxl.readthedocs.io/en/stable/*](https://openpyxl.readthedocs.io/en/stable/)

*дата обращения: (28.04.2025)*