**­­­­­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**

**Черевко М.Е.**

**Преподаватель**

**Дик А.Г.**

**Санкт-Петербург**

**2025 г.**

Оглавление

1. [Цель работы 3](#_Toc198375739)
2. [Описание задачи (формализация задачи) 4](#_Toc198375740)
3. [Теоретическая часть 5](#_Toc198375741)
4. [Основные шаги программы 7](#_Toc198375742)
5. [Блок схема программы 8](#_Toc198375743)
6. [Описание программы 11](#_Toc198375744)
7. [Рекомендации пользователя 13](#_Toc198375745)
8. [Рекомендации программиста 14](#_Toc198375746)
9. [Контрольный пример 15](#_Toc198375747)
10. [Анализ работы 17](#_Toc198375748)
11. [Вывод 47](#_Toc198375749)
12. [Источники 48](#_Toc198375750)

# Цель работы

1. **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: метод Хот-Дек, метод заполнения значением медианы, метод восстановления пропущенного значения сплайн-интерполяцией по присутствующим элементам

# 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:
   * **Хот-Дек** (Hot-Deck Imputation)

Методоснован на поиске похожих строк (доноров) для заполнения пропущенных значений

* + **Заполнение значением медианы** (Median Imputation)

В этом методе пропущенные значения в числовых столбцах заменяются медианой непропущенных значений соответствующего столбца, а в категориальных столбцах — модой. Метод прост в реализации и устойчив к выбросам в числовых данных,.

* + **Сплайн-интерполяция** (Spline Interpolation)

Метод использует сплайн-интерполяцию третьего порядка для заполнения пропусков в числовых и категориальных столбцах

1. **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. **Хот-Дек** (Hot-Deck Imputation)

Метод Хот-Дек основан на поиске похожих строк (доноров) для заполнения пропущенных значений. Датасет кластеризуется с использованием алгоритма MiniBatchKMeans, после чего для каждой строки с пропусками в пределах её кластера применяется алгоритм NearestNeighbors с метрикой Евклидова расстояния, учитывающей пропуски (nan\_euclidean). Пропуски в числовых столбцах, заполняются медианой значений соседей, а в категориальных — модой. Для повышения качества кластеризации числовые столбцы нормализуются, а оставшиеся пропуски заполняются значениями ближайшего соседа из всего датасета, обеспечивая соответствие принципам Хот-Дек. Метод эффективен для сохранения локальных зависимостей в данных, но может быть чувствителен к масштабу числовых признаков.

* 1. **Заполнение значением медианы** (Median Imputation)

В этом методе пропущенные значения в числовых столбцах заменяются медианой непропущенных значений соответствующего столбца, Метод прост в реализации и устойчив к выбросам в числовых данных, что делает его подходящим для столбцов с большими значениями, таких как "Номер банковской карты". Однако он не учитывает локальные зависимости между строками, что может привести к потере информации о структуре данных, особенно при высокой доле пропусков или неравномерном распределении значений.

* 1. **Сплайн-интерполяция** (Spline Interpolation)

Метод использует сплайн-интерполяцию третьего порядка для заполнения пропусков в числовых и категориальных столбцах. Для выполнения интерполяции требуется минимум два непропущенных значения в столбце, чтобы построить сплайн. Если это условие не выполняется, столбец остаётся без изменений. Категориальные столбцы предварительно кодируются с использованием частотного кодирования (доля встречаемости каждой категории), после чего интерполяция применяется к закодированным значениям. После интерполяции категориальные значения раскодируются, выбирая ближайшую категорию по частоте. Сплайн-интерполяция предполагает, что данные имеют некоторую плавную зависимость (например, упорядоченность по индексу, как во временных рядах), поэтому метод не эффективен для данных нерегулярной структурой (например, для признака «Цена дома»).

**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:

Где – исходное значение поля, – восстановленное методом.

**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:
   * **Хот-Дек**
   * **Заполнение значением медианы**
   * **Сплайн-интерполяция**

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

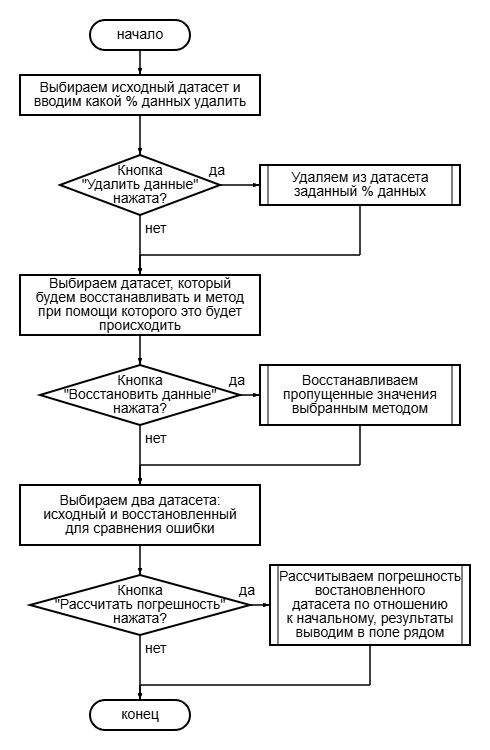
****

Рис. 1 Блок-схема основной программы

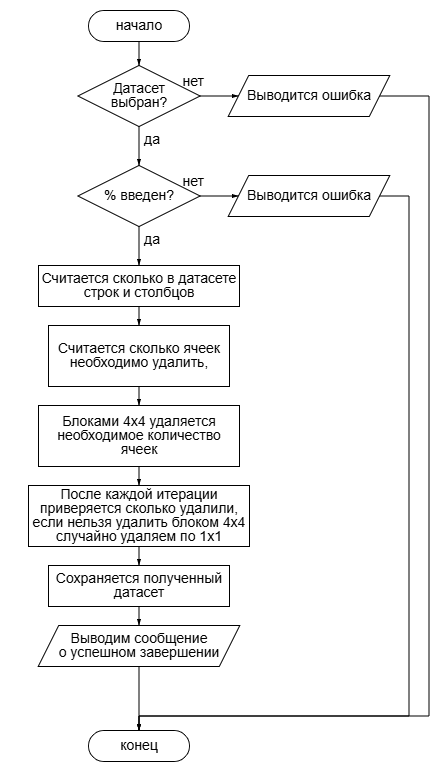
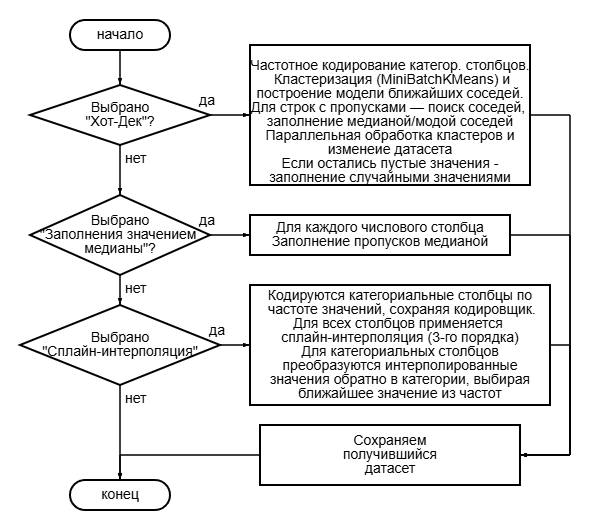


Рис. 2 Блок-схема подпрограммы удаления процента значений



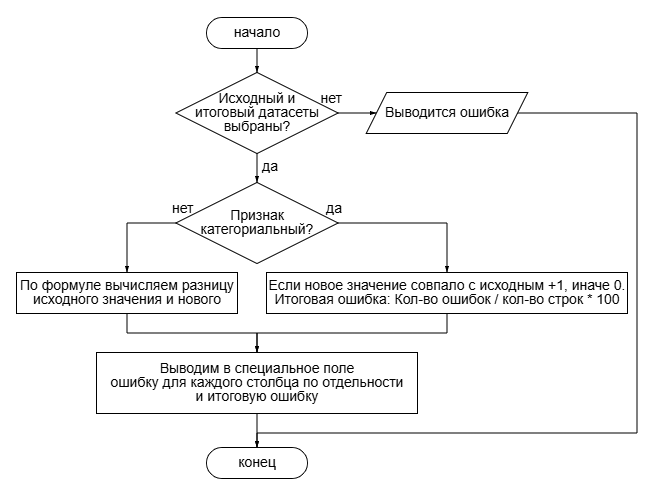
Рис. 3 Блок-схема подпрограммы восстановления пропущенных значений

Рис. 4 Блок-схема подпрограммы подсчета ошибки восстановления

# Описание программы

Программная реализация задачи коммивояжёра написана на языке **Python 3.12.3** с использованием библиотек **tkinter, sklearn, numpy, pandas, joblib, os.** Программа представляет собой графический интерфейс, который позволяет удалять заданный процент 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. Data\_restoration.py

|  |  |  |
| --- | --- | --- |
| Функция | Описание | Возвращаемое значение |
| \_\_init\_\_ | Инициализирует приложение, создаёт окно и переменные для путей файлов. | None |
| setup\_ui | Создаёт интерфейс приложения (панели, кнопки, поля ввода, текстовое поле). | None |
| select\_source | Открывает диалог для выбора исходного Excel-файла и обновляет поле ввода. | None |
| select\_gaps | Открывает диалог для выбора файла с пропусками и обновляет поле ввода. | Non |
| select\_original | Открывает диалог для выбора исходного файла и обновляет поле ввода. | None |
| select\_restored | Открывает диалог для выбора восстановленного файла и обновляет поле ввода. | None |
| update\_entry | Обновляет текстовое поле ввода, вставляя имя файла. | None |
| remove\_data | Удаляет заданный процент данных (блоками 4x4 и одиночными ячейками) из Excel-файла, заменяя их на NaN, и сохраняет результат. | None |
| restore\_data | Восстанавливает пропуски в файле, используя выбранный метод (hot-deck, медиана, сплайн), и сохраняет результат. | None |
| hot\_deck\_imputation | Заполняет пропуски методом hot-deck: кластеризует данные, ищет ближайших соседей и использует их значения (медиана для числовых, мода для категориальных). | DataFrame |
| median\_imputation | Заполняет пропуски в числовых столбцах медианными значениями. | DataFrame |
| spline\_interpolation | Заполняет пропуски сплайн-интерполяцией (3-го порядка) для числовых и категориальных данных, кодируя последние по частоте. | DataFrame |
| calculate\_accuracy | Сравнивает исходный и восстановленный файлы, вычисляет процент ошибок для числовых и категориальных столбцов, выводит результат в текстовое поле. | 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 данные»** — программа создаст копию файла с пропущенными данными (\_nan\_{процент}.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:
   * **Хот-Дек**
   * **Заполнение значением медианы**
   * **Сплайн-интерполяция**
3. Нажмите кнопку **«Восстановить данные»** — будет создан файл с восстановленными значениями (например, \_ restored\_spline.xlsx).

**Оценка качества восстановления:**

1. **Выберите файл ДО удаления данных** — нажмите **«Выбрать»** напротив **«Исходный файл»**.
2. **Выберите восстановленный файл** — нажмите **«Выбрать»** напротив **«Восстановленный файл»**.
3. Нажмите кнопку **«Рассчитать погрешность»**.
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, sklearn, numpy, pandas, joblib, os**. Обычно **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.

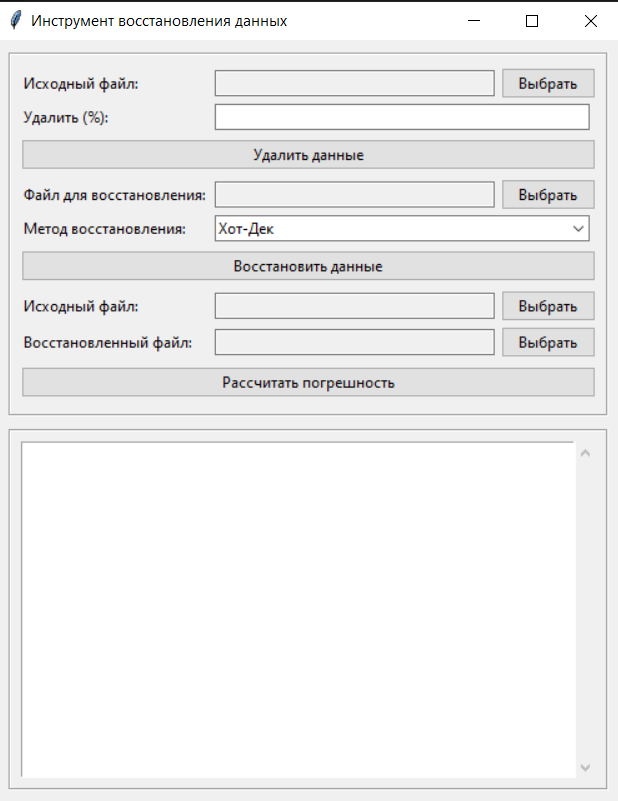
**Код программы:**

<https://github.com/FraaaM/Filling_Dataset_Gaps>

# Контрольный пример

* **Запуск программы:**

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

* **Удаление заданного процента 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)

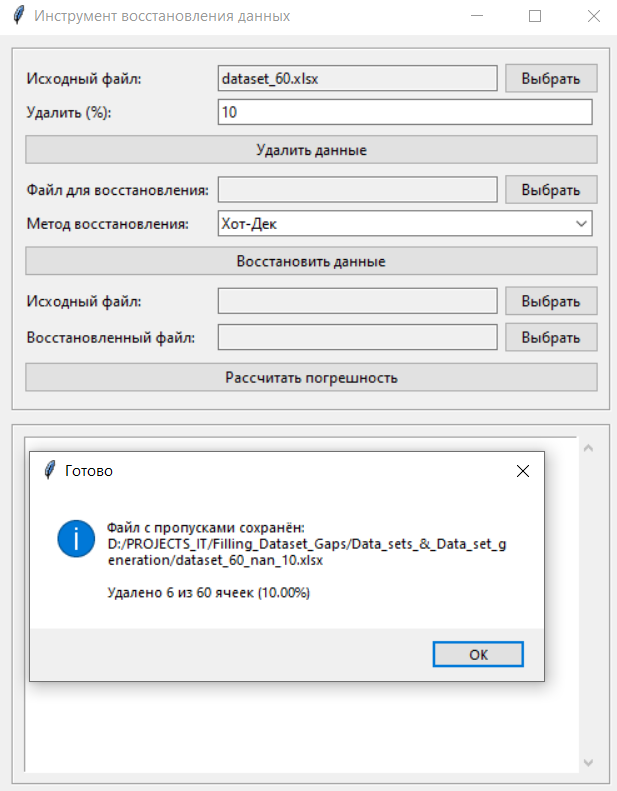


Рис. 5 Графический интерфейс программы Рис. 6 Удаление части данных

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

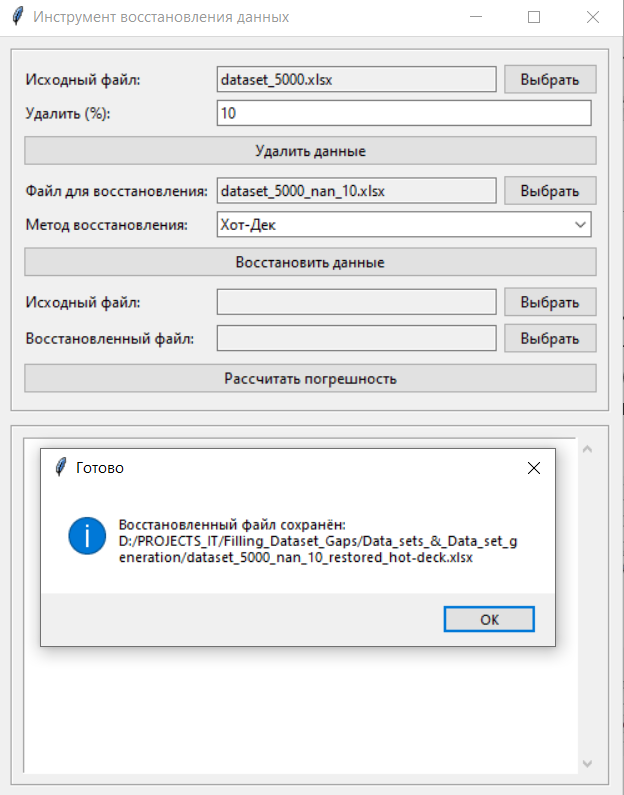
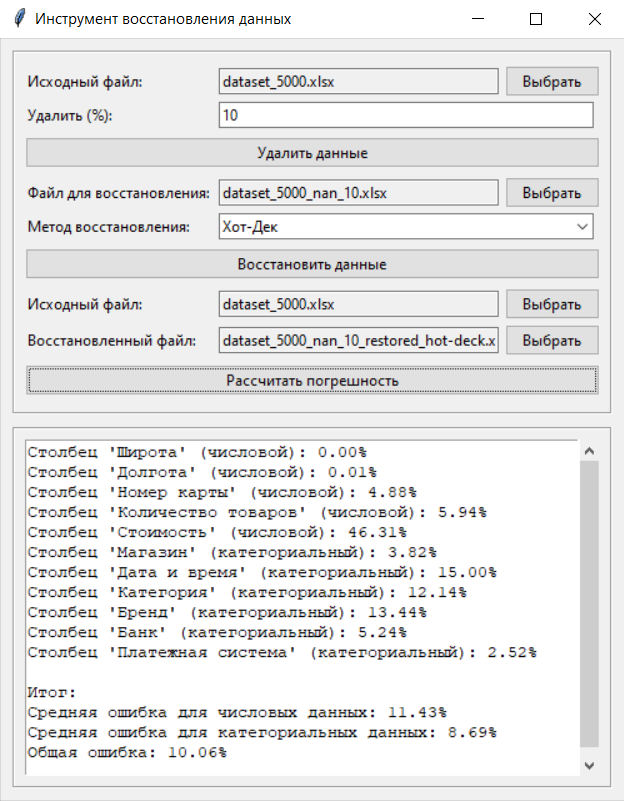


Рис. 7 Восстановление данных Рис. 8 Расчёт погрешности

# Анализ работы

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

Таблица 2. Результаты работы

|  |  |  |  |
| --- | --- | --- | --- |
| Размер датасета | Процент удаления | Метод восстановления | Результаты |
| 5000 | 3 | Хот-Дек | Столбец 'Широта' (числовой): 0.00%  Столбец 'Долгота' (числовой): 0.00%  Столбец 'Номер карты' (числовой): 1.29%  Столбец 'Количество товаров' (числовой): 1.88%  Столбец 'Стоимость' (числовой): 19.01%  Столбец 'Магазин' (категориальный): 1.08%  Столбец 'Дата и время' (категориальный): 4.02%  Столбец 'Категория' (категориальный): 3.34%  Столбец 'Бренд' (категориальный): 3.64%  Столбец 'Банк' (категориальный): 2.06%  Столбец 'Платежная система' (категориальный): 1.04%  Итог:  Средняя ошибка для числовых данных: 4.44%  Средняя ошибка для категориальных данных: 2.53%  Общая ошибка: 3.48% |
| 5000 | 3 | Заполнение значением медианы | Столбец 'Широта' (числовой): 0.00%  Столбец 'Долгота' (числовой): 0.00%  Столбец 'Номер карты' (числовой): 0.94%  Столбец 'Количество товаров' (числовой): 1.88%  Столбец 'Стоимость' (числовой): 16.10%  Столбец 'Магазин' (категориальный): 1.12%  Столбец 'Дата и время' (категориальный): 4.02%  Столбец 'Категория' (категориальный): 4.08%  Столбец 'Бренд' (категориальный): 3.76%  Столбец 'Банк' (категориальный): 4.24%  Столбец 'Платежная система' (категориальный): 3.04%  Итог:  Средняя ошибка для числовых данных: 3.78%  Средняя ошибка для категориальных данных: 3.38%  Общая ошибка: 3.58% |
| 5000 | 3 | Сплайн-интерполяция | Столбец 'Широта' (числовой): 0.00%  Столбец 'Долгота' (числовой): 0.00%  Столбец 'Номер карты' (числовой): 2.88%  Столбец 'Количество товаров' (числовой): 4.33%  Столбец 'Стоимость' (числовой): 81.74%  Столбец 'Магазин' (категориальный): 19.12%  Столбец 'Дата и время' (категориальный): 99.98%  Столбец 'Категория' (категориальный): 19.44%  Столбец 'Бренд' (категориальный): 48.44%  Столбец 'Банк' (категориальный): 2.58%  Столбец 'Платежная система' (категориальный): 1.38%  Итог:  Средняя ошибка для числовых данных: 17.79%  Средняя ошибка для категориальных данных: 31.82%  Общая ошибка: 24.81% |
| 5000 | 5 | Хот-Дек | Столбец 'Широта' (числовой): 0.00%  Столбец 'Долгота' (числовой): 0.00%  Столбец 'Номер карты' (числовой): 2.68%  Столбец 'Количество товаров' (числовой): 3.57%  Столбец 'Стоимость' (числовой): 7.43%  Столбец 'Магазин' (категориальный): 1.60%  Столбец 'Дата и время' (категориальный): 6.74%  Столбец 'Категория' (категориальный): 5.22%  Столбец 'Бренд' (категориальный): 6.94%  Столбец 'Банк' (категориальный): 3.30%  Столбец 'Платежная система' (категориальный): 1.80%  Итог:  Средняя ошибка для числовых данных: 2.74%  Средняя ошибка для категориальных данных: 4.27%  Общая ошибка: 3.50% |
| 5000 | 5 | Заполнение значением медианы | Столбец 'Широта' (числовой): 0.00%  Столбец 'Долгота' (числовой): 0.00%  Столбец 'Номер карты' (числовой): 1.71%  Столбец 'Количество товаров' (числовой): 3.02%  Столбец 'Стоимость' (числовой): 7.26%  Столбец 'Магазин' (категориальный): 1.68%  Столбец 'Дата и время' (категориальный): 6.74%  Столбец 'Категория' (категориальный): 6.64%  Столбец 'Бренд' (категориальный): 6.98%  Столбец 'Банк' (категориальный): 7.00%  Столбец 'Платежная система' (категориальный): 5.36%  Итог:  Средняя ошибка для числовых данных: 2.40%  Средняя ошибка для категориальных данных: 5.73%  Общая ошибка: 4.07% |
| 5000 | 5 | Сплайн-интерполяция | Столбец 'Широта' (числовой): 0.00%  Столбец 'Долгота' (числовой): 0.01%  Столбец 'Номер карты' (числовой): 5.14%  Столбец 'Количество товаров' (числовой): 5.64%  Столбец 'Стоимость' (числовой): 38.96%  Столбец 'Магазин' (категориальный): 32.74%  Столбец 'Дата и время' (категориальный): 99.98%  Столбец 'Категория' (категориальный): 20.94%  Столбец 'Бренд' (категориальный): 49.24%  Столбец 'Банк' (категориальный): 4.68%  Столбец 'Платежная система' (категориальный): 2.38%  Итог:  Средняя ошибка для числовых данных: 9.95%  Средняя ошибка для категориальных данных: 34.99%  Общая ошибка: 22.47% |
| 5000 | 10 | Хот-Дек | Столбец 'Широта' (числовой): 0.00%  Столбец 'Долгота' (числовой): 0.01%  Столбец 'Номер карты' (числовой): 4.88%  Столбец 'Количество товаров' (числовой): 5.94%  Столбец 'Стоимость' (числовой): 46.31%  Столбец 'Магазин' (категориальный): 3.82%  Столбец 'Дата и время' (категориальный): 15.00%  Столбец 'Категория' (категориальный): 12.14%  Столбец 'Бренд' (категориальный): 13.44%  Столбец 'Банк' (категориальный): 5.24%  Столбец 'Платежная система' (категориальный): 2.52%  Итог:  Средняя ошибка для числовых данных: 11.43%  Средняя ошибка для категориальных данных: 8.69%  Общая ошибка: 10.06% |
| 5000 | 10 | Заполнение значением медианы | Столбец 'Широта' (числовой): 0.00%  Столбец 'Долгота' (числовой): 0.01%  Столбец 'Номер карты' (числовой): 2.96%  Столбец 'Количество товаров' (числовой): 5.53%  Столбец 'Стоимость' (числовой): 45.93%  Столбец 'Магазин' (категориальный): 4.00%  Столбец 'Дата и время' (категориальный): 15.00%  Столбец 'Категория' (категориальный): 14.32%  Столбец 'Бренд' (категориальный): 13.56%  Столбец 'Банк' (категориальный): 12.48%  Столбец 'Платежная система' (категориальный): 9.22%  Итог:  Средняя ошибка для числовых данных: 10.89%  Средняя ошибка для категориальных данных: 11.43%  Общая ошибка: 11.16% |
| 5000 | 10 | Сплайн-интерполяция | Столбец 'Широта' (числовой): 0.00%  Столбец 'Долгота' (числовой): 0.01%  Столбец 'Номер карты' (числовой): 9.15%  Столбец 'Количество товаров' (числовой): 10.06%  Столбец 'Стоимость' (числовой): 213.88%  Столбец 'Магазин' (категориальный): 26.00%  Столбец 'Дата и время' (категориальный): 99.98%  Столбец 'Категория' (категориальный): 26.66%  Столбец 'Бренд' (категориальный): 51.68%  Столбец 'Банк' (категориальный): 8.12%  Столбец 'Платежная система' (категориальный): 3.88%  Итог:  Средняя ошибка для числовых данных: 46.62%  Средняя ошибка для категориальных данных: 36.05%  Общая ошибка: 41.34% |
| 5000 | 20 | Хот-Дек | Столбец 'Широта' (числовой): 0.01%  Столбец 'Долгота' (числовой): 0.02%  Столбец 'Номер карты' (числовой): 10.17%  Столбец 'Количество товаров' (числовой): 13.52%  Столбец 'Стоимость' (числовой): 60.66%  Столбец 'Магазин' (категориальный): 6.66%  Столбец 'Дата и время' (категориальный): 28.74%  Столбец 'Категория' (категориальный): 22.58%  Столбец 'Бренд' (категориальный): 23.90%  Столбец 'Банк' (категориальный): 11.44%  Столбец 'Платежная система' (категориальный): 5.12%  Итог:  Средняя ошибка для числовых данных: 16.88%  Средняя ошибка для категориальных данных: 16.41%  Общая ошибка: 16.64% |
| 5000 | 20 | Заполнение значением медианы | Столбец 'Широта' (числовой): 0.01%  Столбец 'Долгота' (числовой): 0.02%  Столбец 'Номер карты' (числовой): 5.27%  Столбец 'Количество товаров' (числовой): 12.88%  Столбец 'Стоимость' (числовой): 59.04%  Столбец 'Магазин' (категориальный): 7.14%  Столбец 'Дата и время' (категориальный): 28.74%  Столбец 'Категория' (категориальный): 27.46%  Столбец 'Бренд' (категориальный): 24.64%  Столбец 'Банк' (категориальный): 26.24%  Столбец 'Платежная система' (категориальный): 20.42%  Итог:  Средняя ошибка для числовых данных: 15.45%  Средняя ошибка для категориальных данных: 22.44%  Общая ошибка: 18.94% |
| 5000 | 20 | Сплайн-интерполяция | Столбец 'Широта' (числовой): 0.01%  Столбец 'Долгота' (числовой): 0.03%  Столбец 'Номер карты' (числовой): 19.30%  Столбец 'Количество товаров' (числовой): 21.04%  Столбец 'Стоимость' (числовой): 249.67%  Столбец 'Магазин' (категориальный): 19.48%  Столбец 'Дата и время' (категориальный): 99.98%  Столбец 'Категория' (категориальный): 38.88%  Столбец 'Бренд' (категориальный): 60.62%  Столбец 'Банк' (категориальный): 17.18%  Столбец 'Платежная система' (категориальный): 10.56%  Итог:  Средняя ошибка для числовых данных: 58.01%  Средняя ошибка для категориальных данных: 41.12%  Общая ошибка: 49.56% |
| 5000 | 30 | Хот-Дек | Столбец 'Широта' (числовой): 0.01%  Столбец 'Долгота' (числовой): 0.03%  Столбец 'Номер карты' (числовой): 13.43%  Столбец 'Количество товаров' (числовой): 21.34%  Столбец 'Стоимость' (числовой): 148.84%  Столбец 'Магазин' (категориальный): 12.86%  Столбец 'Дата и время' (категориальный): 41.60%  Столбец 'Категория' (категориальный): 30.74%  Столбец 'Бренд' (категориальный): 32.78%  Столбец 'Банк' (категориальный): 16.62%  Столбец 'Платежная система' (категориальный): 8.40%  Итог:  Средняя ошибка для числовых данных: 36.73%  Средняя ошибка для категориальных данных: 23.83%  Общая ошибка: 30.28% |
| 5000 | 30 | Заполнение значением медианы | Столбец 'Широта' (числовой): 0.01%  Столбец 'Долгота' (числовой): 0.03%  Столбец 'Номер карты' (числовой): 7.93%  Столбец 'Количество товаров' (числовой): 20.31%  Столбец 'Стоимость' (числовой): 148.29%  Столбец 'Магазин' (категориальный): 13.30%  Столбец 'Дата и время' (категориальный): 41.60%  Столбец 'Категория' (категориальный): 36.50%  Столбец 'Бренд' (категориальный): 33.92%  Столбец 'Банк' (категориальный): 40.88%  Столбец 'Платежная система' (категориальный): 32.74%  Итог:  Средняя ошибка для числовых данных: 35.31%  Средняя ошибка для категориальных данных: 33.16%  Общая ошибка: 34.24% |
| 5000 | 30 | Сплайн-интерполяция | Столбец 'Широта' (числовой): 0.01%  Столбец 'Долгота' (числовой): 0.03%  Столбец 'Номер карты' (числовой): 28.31%  Столбец 'Количество товаров' (числовой): 34.83%  Столбец 'Стоимость' (числовой): 907.72%  Столбец 'Магазин' (категориальный): 25.24%  Столбец 'Дата и время' (категориальный): 99.98%  Столбец 'Категория' (категориальный): 43.36%  Столбец 'Бренд' (категориальный): 68.78%  Столбец 'Банк' (категориальный): 27.44%  Столбец 'Платежная система' (категориальный): 13.10%  Итог:  Средняя ошибка для числовых данных: 194.18%  Средняя ошибка для категориальных данных: 46.32%  Общая ошибка: 120.25% |
| 50000 | 3 | Хот-Дек | Столбец 'Широта' (числовой): 0.00%  Столбец 'Долгота' (числовой): 0.00%  Столбец 'Номер карты' (числовой): 1.46%  Столбец 'Количество товаров' (числовой): 2.00%  Столбец 'Стоимость' (числовой): 10.07%  Столбец 'Магазин' (категориальный): 0.95%  Столбец 'Дата и время' (категориальный): 4.26%  Столбец 'Категория' (категориальный): 3.58%  Столбец 'Бренд' (категориальный): 3.78%  Столбец 'Банк' (категориальный): 1.67%  Столбец 'Платежная система' (категориальный): 0.88%  Итог:  Средняя ошибка для числовых данных: 2.71%  Средняя ошибка для категориальных данных: 2.52%  Общая ошибка: 2.61% |
| 50000 | 3 | Заполнение значением медианы | Столбец 'Широта' (числовой): 0.00%  Столбец 'Долгота' (числовой): 0.00%  Столбец 'Номер карты' (числовой): 0.91%  Столбец 'Количество товаров' (числовой): 2.07%  Столбец 'Стоимость' (числовой): 10.28%  Столбец 'Магазин' (категориальный): 0.99%  Столбец 'Дата и время' (категориальный): 4.26%  Столбец 'Категория' (категориальный): 4.23%  Столбец 'Бренд' (категориальный): 3.90%  Столбец 'Банк' (категориальный): 3.98%  Столбец 'Платежная система' (категориальный): 3.02%  Итог:  Средняя ошибка для числовых данных: 2.65%  Средняя ошибка для категориальных данных: 3.40%  Общая ошибка: 3.03% |
| 50000 | 3 | Сплайн-интерполяция | Столбец 'Широта' (числовой): 0.00%  Столбец 'Долгота' (числовой): 0.00%  Столбец 'Номер карты' (числовой): 2.84%  Столбец 'Количество товаров' (числовой): 3.23%  Столбец 'Стоимость' (числовой): 57.24%  Столбец 'Магазин' (категориальный): 0.94%  Столбец 'Дата и время' (категориальный): 99.99%  Столбец 'Категория' (категориальный): 6.03%  Столбец 'Бренд' (категориальный): 20.29%  Столбец 'Банк' (категориальный): 2.58%  Столбец 'Платежная система' (категориальный): 1.24%  Итог:  Средняя ошибка для числовых данных: 12.66%  Средняя ошибка для категориальных данных: 21.85%  Общая ошибка: 17.25% |
| 50000 | 5 | Хот-Дек | Столбец 'Широта' (числовой): 0.00%  Столбец 'Долгота' (числовой): 0.01%  Столбец 'Номер карты' (числовой): 2.56%  Столбец 'Количество товаров' (числовой): 3.16%  Столбец 'Стоимость' (числовой): 17.55%  Столбец 'Магазин' (категориальный): 1.62%  Столбец 'Дата и время' (категориальный): 7.00%  Столбец 'Категория' (категориальный): 5.88%  Столбец 'Бренд' (категориальный): 6.64%  Столбец 'Банк' (категориальный): 3.30%  Столбец 'Платежная система' (категориальный): 1.31%  Итог:  Средняя ошибка для числовых данных: 4.65%  Средняя ошибка для категориальных данных: 4.29%  Общая ошибка: 4.47% |
| 50000 | 5 | Заполнение значением медианы | Столбец 'Широта' (числовой): 0.00%  Столбец 'Долгота' (числовой): 0.01%  Столбец 'Номер карты' (числовой): 1.45%  Столбец 'Количество товаров' (числовой): 3.26%  Столбец 'Стоимость' (числовой): 17.22%  Столбец 'Магазин' (категориальный): 1.68%  Столбец 'Дата и время' (категориальный): 7.00%  Столбец 'Категория' (категориальный): 7.04%  Столбец 'Бренд' (категориальный): 6.78%  Столбец 'Банк' (категориальный): 6.75%  Столбец 'Платежная система' (категориальный): 5.04%  Итог:  Средняя ошибка для числовых данных: 4.39%  Средняя ошибка для категориальных данных: 5.71%  Общая ошибка: 5.05% |
| 50000 | 5 | Сплайн-интерполяция | Столбец 'Широта' (числовой): 0.00%  Столбец 'Долгота' (числовой): 0.01%  Столбец 'Номер карты' (числовой): 4.51%  Столбец 'Количество товаров' (числовой): 5.06%  Столбец 'Стоимость' (числовой): 87.19%  Столбец 'Магазин' (категориальный): 1.59%  Столбец 'Дата и время' (категориальный): 99.99%  Столбец 'Категория' (категориальный): 7.94%  Столбец 'Бренд' (категориальный): 25.63%  Столбец 'Банк' (категориальный): 4.45%  Столбец 'Платежная система' (категориальный): 1.98%  Итог:  Средняя ошибка для числовых данных: 19.35%  Средняя ошибка для категориальных данных: 23.60%  Общая ошибка: 21.48% |
| 50000 | 10 | Хот-Дек | Столбец 'Широта' (числовой): 0.00%  Столбец 'Долгота' (числовой): 0.01%  Столбец 'Номер карты' (числовой): 13.67%  Столбец 'Количество товаров' (числовой): 6.49%  Столбец 'Стоимость' (числовой): 36.28%  Столбец 'Магазин' (категориальный): 3.40%  Столбец 'Дата и время' (категориальный): 13.61%  Столбец 'Категория' (категориальный): 11.17%  Столбец 'Бренд' (категориальный): 12.82%  Столбец 'Банк' (категориальный): 7.04%  Столбец 'Платежная система' (категориальный): 2.77%  Итог:  Средняя ошибка для числовых данных: 11.29%  Средняя ошибка для категориальных данных: 8.47%  Общая ошибка: 9.88% |
| 50000 | 10 | Заполнение значением медианы | Столбец 'Широта' (числовой): 0.00%  Столбец 'Долгота' (числовой): 0.01%  Столбец 'Номер карты' (числовой): 2.92%  Столбец 'Количество товаров' (числовой): 7.46%  Столбец 'Стоимость' (числовой): 35.67%  Столбец 'Магазин' (категориальный): 3.54%  Столбец 'Дата и время' (категориальный): 13.61%  Столбец 'Категория' (категориальный): 13.34%  Столбец 'Бренд' (категориальный): 13.02%  Столбец 'Банк' (категориальный): 13.89%  Столбец 'Платежная система' (категориальный): 10.62%  Итог:  Средняя ошибка для числовых данных: 9.21%  Средняя ошибка для категориальных данных: 11.34%  Общая ошибка: 10.28% |
| 50000 | 10 | Сплайн-интерполяция | Столбец 'Широта' (числовой): 0.00%  Столбец 'Долгота' (числовой): 0.01%  Столбец 'Номер карты' (числовой): 9.78%  Столбец 'Количество товаров' (числовой): 10.74%  Столбец 'Стоимость' (числовой): 183.08%  Столбец 'Магазин' (категориальный): 3.35%  Столбец 'Дата и время' (категориальный): 99.99%  Столбец 'Категория' (категориальный): 15.14%  Столбец 'Бренд' (категориальный): 31.48%  Столбец 'Банк' (категориальный): 9.25%  Столбец 'Платежная система' (категориальный): 4.20%  Итог:  Средняя ошибка для числовых данных: 40.72%  Средняя ошибка для категориальных данных: 27.23%  Общая ошибка: 33.98% |
| 50000 | 20 | Хот-Дек | Столбец 'Широта' (числовой): 0.01%  Столбец 'Долгота' (числовой): 0.02%  Столбец 'Номер карты' (числовой): 9.87%  Столбец 'Количество товаров' (числовой): 12.63%  Столбец 'Стоимость' (числовой): 71.18%  Столбец 'Магазин' (категориальный): 7.28%  Столбец 'Дата и время' (категориальный): 27.81%  Столбец 'Категория' (категориальный): 22.61%  Столбец 'Бренд' (категориальный): 24.61%  Столбец 'Банк' (категориальный): 11.21%  Столбец 'Платежная система' (категориальный): 5.71%  Итог:  Средняя ошибка для числовых данных: 18.74%  Средняя ошибка для категориальных данных: 16.54%  Общая ошибка: 17.64% |
| 50000 | 20 | Заполнение значением медианы | Столбец 'Широта' (числовой): 0.01%  Столбец 'Долгота' (числовой): 0.02%  Столбец 'Номер карты' (числовой): 5.76%  Столбец 'Количество товаров' (числовой): 14.49%  Столбец 'Стоимость' (числовой): 69.26%  Столбец 'Магазин' (категориальный): 7.62%  Столбец 'Дата и время' (категориальный): 27.81%  Столбец 'Категория' (категориальный): 26.20%  Столбец 'Бренд' (категориальный): 25.38%  Столбец 'Банк' (категориальный): 27.19%  Столбец 'Платежная система' (категориальный): 21.18%  Итог:  Средняя ошибка для числовых данных: 17.91%  Средняя ошибка для категориальных данных: 22.56%  Общая ошибка: 20.24% |
| 50000 | 20 | Сплайн-интерполяция | Столбец 'Широта' (числовой): 0.01%  Столбец 'Долгота' (числовой): 0.02%  Столбец 'Номер карты' (числовой): 19.37%  Столбец 'Количество товаров' (числовой): 20.52%  Столбец 'Стоимость' (числовой): 382.63%  Столбец 'Магазин' (категориальный): 11.46%  Столбец 'Дата и время' (категориальный): 99.99%  Столбец 'Категория' (категориальный): 29.44%  Столбец 'Бренд' (категориальный): 38.88%  Столбец 'Банк' (категориальный): 17.89%  Столбец 'Платежная система' (категориальный): 8.44%  Итог:  Средняя ошибка для числовых данных: 84.51%  Средняя ошибка для категориальных данных: 34.35%  Общая ошибка: 59.43% |
| 50000 | 30 | Хот-Дек | Столбец 'Широта' (числовой): 0.01%  Столбец 'Долгота' (числовой): 0.03%  Столбец 'Номер карты' (числовой): 14.15%  Столбец 'Количество товаров' (числовой): 20.31%  Столбец 'Стоимость' (числовой): 119.36%  Столбец 'Магазин' (категориальный): 11.43%  Столбец 'Дата и время' (категориальный): 40.99%  Столбец 'Категория' (категориальный): 31.23%  Столбец 'Бренд' (категориальный): 34.85%  Столбец 'Банк' (категориальный): 16.59%  Столбец 'Платежная система' (категориальный): 8.82%  Итог:  Средняя ошибка для числовых данных: 30.77%  Средняя ошибка для категориальных данных: 23.99%  Общая ошибка: 27.38% |
| 50000 | 30 | Заполнение значением медианы | Столбец 'Широта' (числовой): 0.01%  Столбец 'Долгота' (числовой): 0.03%  Столбец 'Номер карты' (числовой): 8.16%  Столбец 'Количество товаров' (числовой): 23.21%  Столбец 'Стоимость' (числовой): 117.79%  Столбец 'Магазин' (категориальный): 11.97%  Столбец 'Дата и время' (категориальный): 40.99%  Столбец 'Категория' (категориальный): 37.40%  Столбец 'Бренд' (категориальный): 35.90%  Столбец 'Банк' (категориальный): 41.51%  Столбец 'Платежная система' (категориальный): 33.13%  Итог:  Средняя ошибка для числовых данных: 29.84%  Средняя ошибка для категориальных данных: 33.48%  Общая ошибка: 31.66% |
| 50000 | 30 | Сплайн-интерполяция | Столбец 'Широта' (числовой): 0.01%  Столбец 'Долгота' (числовой): 0.05%  Столбец 'Номер карты' (числовой): 28.07%  Столбец 'Количество товаров' (числовой): 34.11%  Столбец 'Стоимость' (числовой): 662.77%  Столбец 'Магазин' (категориальный): 11.45%  Столбец 'Дата и время' (категориальный): 99.99%  Столбец 'Категория' (категориальный): 38.01%  Столбец 'Бренд' (категориальный): 50.49%  Столбец 'Банк' (категориальный): 27.39%  Столбец 'Платежная система' (категориальный): 13.29%  Итог:  Средняя ошибка для числовых данных: 145.00%  Средняя ошибка для категориальных данных: 40.10%  Общая ошибка: 92.55% |
| 120000 | 3 | Хот-Дек | Столбец 'Широта' (числовой): 0.00%  Столбец 'Долгота' (числовой): 0.00%  Столбец 'Номер карты' (числовой): 1.49%  Столбец 'Количество товаров' (числовой): 1.85%  Столбец 'Стоимость' (числовой): 12.19%  Столбец 'Магазин' (категориальный): 1.05%  Столбец 'Дата и время' (категориальный): 4.19%  Столбец 'Категория' (категориальный): 3.40%  Столбец 'Бренд' (категориальный): 3.93%  Столбец 'Банк' (категориальный): 1.98%  Столбец 'Платежная система' (категориальный): 0.76%  Итог:  Средняя ошибка для числовых данных: 3.11%  Средняя ошибка для категориальных данных: 2.55%  Общая ошибка: 2.83% |
| 120000 | 3 | Заполнение значением медианы | Столбец 'Широта' (числовой): 0.00%  Столбец 'Долгота' (числовой): 0.00%  Столбец 'Номер карты' (числовой): 0.85%  Столбец 'Количество товаров' (числовой): 2.10%  Столбец 'Стоимость' (числовой): 12.24%  Столбец 'Магазин' (категориальный): 1.11%  Столбец 'Дата и время' (категориальный): 4.19%  Столбец 'Категория' (категориальный): 4.08%  Столбец 'Бренд' (категориальный): 4.03%  Столбец 'Банк' (категориальный): 4.06%  Столбец 'Платежная система' (категориальный): 3.06%  Итог:  Средняя ошибка для числовых данных: 3.04%  Средняя ошибка для категориальных данных: 3.42%  Общая ошибка: 3.23% |
| 120000 | 3 | Сплайн-интерполяция | Столбец 'Широта' (числовой): 0.00%  Столбец 'Долгота' (числовой): 0.00%  Столбец 'Номер карты' (числовой): 2.69%  Столбец 'Количество товаров' (числовой): 3.17%  Столбец 'Стоимость' (числовой): 65.66%  Столбец 'Магазин' (категориальный): 5.55%  Столбец 'Дата и время' (категориальный): 100.00%  Столбец 'Категория' (категориальный): 4.50%  Столбец 'Бренд' (категориальный): 15.30%  Столбец 'Банк' (категориальный): 2.70%  Столбец 'Платежная система' (категориальный): 1.21%  Итог:  Средняя ошибка для числовых данных: 14.30%  Средняя ошибка для категориальных данных: 21.54%  Общая ошибка: 17.92% |
| 120000 | 5 | Хот-Дек | Столбец 'Широта' (числовой): 0.00%  Столбец 'Долгота' (числовой): 0.00%  Столбец 'Номер карты' (числовой): 1.56%  Столбец 'Количество товаров' (числовой): 3.37%  Столбец 'Стоимость' (числовой): 17.33%  Столбец 'Магазин' (категориальный): 1.66%  Столбец 'Дата и время' (категориальный): 6.73%  Столбец 'Категория' (категориальный): 5.54%  Столбец 'Бренд' (категориальный): 6.66%  Столбец 'Банк' (категориальный): 2.84%  Столбец 'Платежная система' (категориальный): 1.47%  Итог:  Средняя ошибка для числовых данных: 4.45%  Средняя ошибка для категориальных данных: 4.15%  Общая ошибка: 4.30% |
| 120000 | 5 | Заполнение значением медианы | Столбец 'Широта' (числовой): 0.00%  Столбец 'Долгота' (числовой): 0.00%  Столбец 'Номер карты' (числовой): 1.56%  Столбец 'Количество товаров' (числовой): 3.49%  Столбец 'Стоимость' (числовой): 17.19%  Столбец 'Магазин' (категориальный): 1.75%  Столбец 'Дата и время' (категориальный): 6.73%  Столбец 'Категория' (категориальный): 6.64%  Столбец 'Бренд' (категориальный): 6.88%  Столбец 'Банк' (категориальный): 7.02%  Столбец 'Платежная система' (категориальный): 5.36%  Итог:  Средняя ошибка для числовых данных: 4.45%  Средняя ошибка для категориальных данных: 5.73%  Общая ошибка: 5.09% |
| 120000 | 5 | Сплайн-интерполяция | Столбец 'Широта' (числовой): 0.00%  Столбец 'Долгота' (числовой): 0.01%  Столбец 'Номер карты' (числовой): 5.04%  Столбец 'Количество товаров' (числовой): 5.01%  Столбец 'Стоимость' (числовой): 93.53%  Столбец 'Магазин' (категориальный): 10.61%  Столбец 'Дата и время' (категориальный): 100.00%  Столбец 'Категория' (категориальный): 7.65%  Столбец 'Бренд' (категориальный): 17.16%  Столбец 'Банк' (категориальный): 4.64%  Столбец 'Платежная система' (категориальный): 2.16%  Итог:  Средняя ошибка для числовых данных: 20.72%  Средняя ошибка для категориальных данных: 23.70%  Общая ошибка: 22.21% |
| 120000 | 10 | Хот-Дек | Столбец 'Широта' (числовой): 0.00%  Столбец 'Долгота' (числовой): 0.01%  Столбец 'Номер карты' (числовой): 5.02%  Столбец 'Количество товаров' (числовой): 6.02%  Столбец 'Стоимость' (числовой): 36.82%  Столбец 'Магазин' (категориальный): 3.54%  Столбец 'Дата и время' (категориальный): 13.73%  Столбец 'Категория' (категориальный): 11.06%  Столбец 'Бренд' (категориальный): 12.94%  Столбец 'Банк' (категориальный): 6.87%  Столбец 'Платежная система' (категориальный): 2.72%  Итог:  Средняя ошибка для числовых данных: 9.58%  Средняя ошибка для категориальных данных: 8.48%  Общая ошибка: 9.03% |
| 120000 | 10 | Заполнение значением медианы | Столбец 'Широта' (числовой): 0.00%  Столбец 'Долгота' (числовой): 0.01%  Столбец 'Номер карты' (числовой): 2.90%  Столбец 'Количество товаров' (числовой): 6.84%  Столбец 'Стоимость' (числовой): 36.57%  Столбец 'Магазин' (категориальный): 3.70%  Столбец 'Дата и время' (категориальный): 13.73%  Столбец 'Категория' (категориальный): 13.38%  Столбец 'Бренд' (категориальный): 13.33%  Столбец 'Банк' (категориальный): 13.77%  Столбец 'Платежная система' (категориальный): 10.43%  Итог:  Средняя ошибка для числовых данных: 9.26%  Средняя ошибка для категориальных данных: 11.39%  Общая ошибка: 10.33% |
| 120000 | 10 | Сплайн-интерполяция | Столбец 'Широта' (числовой): 0.00%  Столбец 'Долгота' (числовой): 0.01%  Столбец 'Номер карты' (числовой): 9.57%  Столбец 'Количество товаров' (числовой): 10.11%  Столбец 'Стоимость' (числовой): 181.54%  Столбец 'Магазин' (категориальный): 12.28%  Столбец 'Дата и время' (категориальный): 100.00%  Столбец 'Категория' (категориальный): 15.45%  Столбец 'Бренд' (категориальный): 23.25%  Столбец 'Банк' (категориальный): 9.14%  Столбец 'Платежная система' (категориальный): 4.10%  Итог:  Средняя ошибка для числовых данных: 40.25%  Средняя ошибка для категориальных данных: 27.37%  Общая ошибка: 33.81% |
| 120000 | 20 | Хот-Дек | Столбец 'Широта' (числовой): 0.01%  Столбец 'Долгота' (числовой): 0.02%  Столбец 'Номер карты' (числовой): 9.86%  Столбец 'Количество товаров' (числовой): 14.10%  Столбец 'Стоимость' (числовой): 74.93%  Столбец 'Магазин' (категориальный): 7.34%  Столбец 'Дата и время' (категориальный): 27.77%  Столбец 'Категория' (категориальный): 21.75%  Столбец 'Бренд' (категориальный): 25.04%  Столбец 'Банк' (категориальный): 10.83%  Столбец 'Платежная система' (категориальный): 5.61%  Итог:  Средняя ошибка для числовых данных: 19.78%  Средняя ошибка для категориальных данных: 16.39%  Общая ошибка: 18.09% |
| 120000 | 20 | Заполнение значением медианы | Столбец 'Широта' (числовой): 0.01%  Столбец 'Долгота' (числовой): 0.02%  Столбец 'Номер карты' (числовой): 5.68%  Столбец 'Количество товаров' (числовой): 14.68%  Столбец 'Стоимость' (числовой): 74.92%  Столбец 'Магазин' (категориальный): 7.68%  Столбец 'Дата и время' (категориальный): 27.77%  Столбец 'Категория' (категориальный): 26.17%  Столбец 'Бренд' (категориальный): 25.61%  Столбец 'Банк' (категориальный): 27.23%  Столбец 'Платежная система' (категориальный): 21.15%  Итог:  Средняя ошибка для числовых данных: 19.06%  Средняя ошибка для категориальных данных: 22.60%  Общая ошибка: 20.83% |
| 120000 | 20 | Сплайн-интерполяция | Столбец 'Широта' (числовой): 0.01%  Столбец 'Долгота' (числовой): 0.02%  Столбец 'Номер карты' (числовой): 18.88%  Столбец 'Количество товаров' (числовой): 22.16%  Столбец 'Стоимость' (числовой): 411.76%  Столбец 'Магазин' (категориальный): 7.32%  Столбец 'Дата и время' (категориальный): 100.00%  Столбец 'Категория' (категориальный): 26.46%  Столбец 'Бренд' (категориальный): 34.45%  Столбец 'Банк' (категориальный): 17.96%  Столбец 'Платежная система' (категориальный): 8.43%  Итог:  Средняя ошибка для числовых данных: 90.57%  Средняя ошибка для категориальных данных: 32.44%  Общая ошибка: 61.50% |
| 120000 | 30 | Хот-Дек | Столбец 'Широта' (числовой): 0.01%  Столбец 'Долгота' (числовой): 0.03%  Столбец 'Номер карты' (числовой): 14.04%  Столбец 'Количество товаров' (числовой): 22.25%  Столбец 'Стоимость' (числовой): 130.14%  Столбец 'Магазин' (категориальный): 11.62%  Столбец 'Дата и время' (категориальный): 41.41%  Столбец 'Категория' (категориальный): 31.40%  Столбец 'Бренд' (категориальный): 34.91%  Столбец 'Банк' (категориальный): 16.18%  Столбец 'Платежная система' (категориальный): 8.71%  Итог:  Средняя ошибка для числовых данных: 33.30%  Средняя ошибка для категориальных данных: 24.04%  Общая ошибка: 28.67% |
| 120000 | 30 | Заполнение значением медианы | Столбец 'Широта' (числовой): 0.01%  Столбец 'Долгота' (числовой): 0.03%  Столбец 'Номер карты' (числовой): 8.06%  Столбец 'Количество товаров' (числовой): 23.08%  Столбец 'Стоимость' (числовой): 128.65%  Столбец 'Магазин' (категориальный): 12.14%  Столбец 'Дата и время' (категориальный): 41.41%  Столбец 'Категория' (категориальный): 37.72%  Столбец 'Бренд' (категориальный): 36.05%  Столбец 'Банк' (категориальный): 41.09%  Столбец 'Платежная система' (категориальный): 32.64%  Итог:  Средняя ошибка для числовых данных: 31.97%  Средняя ошибка для категориальных данных: 33.51%  Общая ошибка: 32.74% |
| 120000 | 30 | Сплайн-интерполяция | Столбец 'Широта' (числовой): 0.01%  Столбец 'Долгота' (числовой): 0.03%  Столбец 'Номер карты' (числовой): 27.83%  Столбец 'Количество товаров' (числовой): 34.64%  Столбец 'Стоимость' (числовой): 677.83%  Столбец 'Магазин' (категориальный): 15.53%  Столбец 'Дата и время' (категориальный): 100.00%  Столбец 'Категория' (категориальный): 37.53%  Столбец 'Бренд' (категориальный): 45.20%  Столбец 'Банк' (категориальный): 27.24%  Столбец 'Платежная система' (категориальный): 13.02%  Итог:  Средняя ошибка для числовых данных: 148.07%  Средняя ошибка для категориальных данных: 39.75%  Общая ошибка: 93.91% |

На основе представленной таблицы результатов проведено сравнение трёх методов восстановления данных: Хот-Дек, Заполнение значением медианы и Сплайн-интерполяция. Анализ выполнен с учётом размера датасета (5 000, 12 000, 50 000, 120 000 строк), процента удаления данных (3–30%) и типа данных (числовые/категориальные).

**Ключевые наблюдения:**

**1. Хот-Дек:**

Преимущества:

* Наименьшая общая ошибка при небольших процентах удаления (3–10%). Например, для датасета 5 000 строк и 3% удаления общая ошибка составляет 3.48%.
* Стабильная работа с категориальными данными. Например, для столбца 'Магазин' ошибка редко превышает 5% даже при 30% удаления.

Недостатки:

* При высоких процентах удаления (20–30%) ошибка возрастает, особенно для числовых столбцов (например, 'Стоимость' достигает 148.84% при 30% удаления).
* Средняя ошибка для числовых данных растёт быстрее, чем для категориальных.

**2. Заполнение значением медианы:**

Преимущества:

* Умеренные ошибки для числовых данных. Например, для датасета 12 000 строк и 5% удаления средняя ошибка числовых данных — 4.45%.
* Лучше справляется с восстановлением категориальных данных по сравнению со Сплайн-интерполяцией (например, ошибка 'Банк' при 20% удаления — 27.19% против 41.12% у сплайна).

Недостатки:

* Неэффективен для категориальных данных при высоком проценте удаления (ошибка 'Платежная система' достигает 32.74% при 30% удаления).
* Общая ошибка растёт с увеличением процента удаления, но медленнее, чем у сплайна.

3. Сплайн-интерполяция:

Преимущества:

* Подходит для восстановления данных с низкой дисперсией (например, 'Широта' и 'Долгота' с ошибкой 0.00–0.05%).

Недостатки:

* Катастрофически высокая ошибка для числовых данных при больших пропусках. Например, для 'Стоимости' при 30% удаления ошибка достигает 907.72% (датасет 5 000 строк).
* Плохая работа с категориальными данными (ошибка 'Дата и время' — 99.98–100% во всех случаях).

Влияние размера датасета:

* На больших датасетах (например, 120 000 строк) методы демонстрируют схожую динамику ошибок, но с меньшим разбросом. Например, для Хот-Дек при 30% удаления общая ошибка составляет 28.67% против 31.66% у заполнения медианой.
* Сплайн-интерполяция остаётся наименее надёжным методом независимо от размера данных.

Выводы:

1. Хот-Дек — оптимальный выбор для:

* Небольших процентов удаления (3–10%).
* Работы с категориальными данными.
* Ситуаций, где критична стабильность результатов.
* Также хорошо себя показывает и при высоком проценте пропусков.

2. Заполнение медианой — компромиссный вариант:

* Для числовых данных с умеренными пропусками (до 20%).
* Когда требуется баланс между простотой реализации и точностью.

3. Сплайн-интерполяция:

* Рекомендуется только для данных с высокой корреляцией и минимальными пропусками (например, геолокационные данные).
* Не подходит для категориальных данных и случаев с высоким процентом удаления.
* применять с осторожностью, только после анализа структуры данных.

# Вывод

В рамках данной работы были изучены и реализованы три метода заполнения пропусков в датасетах: Хот-Дек, заполнение значением медианы и Сплайн-интерполяция. Проведено сравнительное тестирование этих методов на датасетах различного размера (5 000, 50 000 и 120 000 строк) с разным процентом удалённых данных (3%, 5%, 10%, 20%, 30%).Основные результаты:

* **Хот-Дек** оптимален для небольших пропусков (3–10%), работы с категориальными данными и ситуаций, где критична стабильность. Несмотря на рост ошибок при высоких пропусках, сохраняет относительную устойчивость.
* **Заполнение значением медианы** -подходит для числовых данных с умеренными пропусками (до 20%) и случаев, где важен баланс между простотой и точностью.
* **Сплайн-интерполяция** применима только для специфических задач (например, геолокационные данные с низкой дисперсией) и требует тщательного анализа структуры данных перед использованием.

Таким образом, Хот-Дек является предпочтительным методом для большинства случаев.

# Источники

* + *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%20)
    - *Дата обращения: 15.05.2025*
  + *pandas — Библиотека для анализа и обработки данных.* 
    - [*https://pandas.pydata.org/docs/*](https://pandas.pydata.org/docs/)
    - *Дата обращения: 15.05.2025*
  + *scikit-learn — Библиотека для машинного обучения (кластеризация, поиск ближайших соседей).* 
    - [*https://scikit-learn.org/stable/*](https://scikit-learn.org/stable/%20%20)
    - *Дата обращения: 15.05.2025*
  + *numpy — Библиотека для работы с многомерными массивами и математическими операциями.* 
    - [*https://numpy.org/doc/*](https://numpy.org/doc/%20%20)
    - *Дата обращения: 15.05.2025*
  + *joblib — Инструмент для параллельных вычислений в Python.* 
    - [*https://joblib.readthedocs.io/en/latest/*](https://joblib.readthedocs.io/en/latest/%20%20)
    - *Дата обращения: 15.05.2025*
  + *openpyxl — Библиотека для чтения и записи файлов Excel.* 
    - [*https://openpyxl.readthedocs.io/en/stable/*](https://openpyxl.readthedocs.io/en/stable/%20%20)
    - *Дата обращения: 15.05.2025*
  + *scipy — Библиотека для научных вычислений (используется в сплайн-интерполяции).* 
    - [*https://docs.scipy.org/doc/scipy/*](https://docs.scipy.org/doc/scipy/%20%20)
    - *Дата обращения: 15.05.2025*