**САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ**

**Факультет прикладной математики-процессов управления**

**Программа бакалавриата**

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

**ОТЧЕТ**

**по лабораторной работе №4**

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

**на тему «Методы заполнения пропусков»**

**Студент гр. 23Б15-пу**

**Голуб Г.Я.**

**Преподаватель**

**Дик А.Г.**

**Санкт-Петербург**

**2025 г.**

**Оглавление**

[Цель работы 4](#_8h4bt0qrnpil)

[Описание задачи (формализация задачи) 4](#_jr7tniqipiq6)

[Теоретическая часть 5](#_yu58wngm0rao)

[Основные шаги программы 5](#_unzzj6dpxoyq)

[Блок схема программы 6](#_g5ubl0xtzrq4)

[Описание программы 8](#_p7zkqw6z5n9d)

[Рекомендации пользователя 9](#_x6yhnxdul2t5)

[Рекомендации программиста 9](#_ua336n1pg5fr)

[Исходный код программы 9](#_ylnn3yxsvg4h)

[Контрольный пример 9](#_b7iqb9aza4zp)

[Сравнение методов восстановления 11](#_mfxtew9i6zh4)

[Вывод 11](#_fgjfpjnhl0xs)

[Источники 13](#_6mw6qcjz1j89)

# Цель работы

Целью данной лабораторной работы является разработка и реализация программного средства для восстановления пропущенных значений в табличных данных с использованием трёх методов — Hot-deck, LOCF (Last Observation Carried Forward) и сплайн-интерполяции по присутствующим элементам — а также проведение их экспериментального сравнения по метрикам суммарной относительной ошибки и доли некорректных восстановлений на синтетических датасетах различных размеров и уровней пропусков.

# Описание задачи (формализация задачи)

В данной работе рассматривается табличный датасет, содержащий как числовые, так и категориальные признаки, в котором часть ячеек случайным образом заменена на пропуски (NaN) в заранее заданных долях (3 %, 5 %, 10 %, 20 %, 30 %).

Задача состоит в том, чтобы для каждого такого «прореженного» датасета восстановить недостающие значения тремя строго заданными методами и затем оценить качество восстановления.  
Критерии оценки:

1. Для числовых признаков качество измеряется суммарной относительной ошибкой между истинными и восстановленными значениями.
2. Для категориальных — долей неправильных восстановлений среди всех заполненных пропусков.

Эксперимент проводится на трёх синтетических датасетах разного размера и при различных уровнях пропусков. По полученным метрикам суммарной ошибки и доли ошибок делается вывод о том, какой из методов более точен и применим в задаче восстановления реальных медицинских данных.

# Теоретическая часть

**Метод 1. Стратифицированный Hot-deck** Каждое пропущенное значение в столбце заполняется случайной выборкой из «донорского» пула существующих (не NaN) значений того же столбца, причём пул ограничивается внутри заранее определённой когортной группы:

1. Для полей, связанных с визитом (симптомы, врач, даты, анализы, цена), донор выбирается среди записей одного и того же врача;
2. Для паспортных и идентификационных полей (ФИО, паспортные данные, СНИЛС) — среди записей одной и той же страны;  
   Во всех остальных случаях — из всего столбца.

**Метод 2. LOCF (Last Observation Carried Forward) по пациенту** Пропуски заполняются последним известным значением предыдущей строки, а при отсутствии предыдущего — первым доступным последующим, но только в пределах одного пациента (группировка по ФИО). Сначала выполняется «прямое» заполнение (ffill), затем обратное (bfill), что обеспечивает отсутствие NaN внутри всех записей одного и того же пациента.

**Метод 3. Сплайн-интерполяция (для числовых признаков)** Для каждого числового столбца строится гладкая интерполяционная кривая (обычно кубический сплайн) на основе известных точек (не NaN). В позициях пропусков вычисляются значения сплайна, что обеспечивает учёт и предшествующих, и последующих наблюдений и корректное восстановление трендов. Категориальные столбцы данным методом не обрабатываются.

# Основные шаги программы

1. Выбрать датасет на котором будет проводиться эксперимент (маленький, средний, большой).
2. Испортить датасет, добавив пропуски в него.
3. Восстановить датасет, используя выше приведенные методы.
4. Вывести таблицу сравнения алгоритмов

# Описание программы

Программа реализована на языке Python с использованием библиотек PyQt5 [[1]](#_6mw6qcjz1j89) для построения графического интерфейса, pandas [[2]](#_6mw6qcjz1j89) и NumPy [[3]](#_6mw6qcjz1j89) для загрузки, манипуляции и анализа табличных данных, а также SciPy [[4]](#_6mw6qcjz1j89) (через метод spline в pandas) для сплайн-интерполяции числовых пропусков.

Таблица 1. main.py

| Функция | Описание | Возвращаемое значение |
| --- | --- | --- |
| introduce\_missing\_df | Портит датасет | None |
| impute\_hot\_deck\_strat | Применяет метод Хот-Дек | DataFrame |
| impute\_spline | Применяет метод сплайн-интерполяции | DataFrame |
| impute\_locf\_by\_patient | Применяет метод LOCF | DataFrame |
| evaluate\_imputation | Оценивает качество восстановления | DataFrame |
| impute\_dataset | Главная функция, запускающая восстановление датасета | None |
| degrade\_dataset | Запускает удаление ячеек из датасета | None |

# Рекомендации пользователя

* Убедитесь, что в папке с приложением присутствуют три файла: small.csv, medium.csv и large.csv, с одинаковой структурой столбцов.
* Выберите нужный датасет в выпадающем списке.
* Задайте процент пропусков и нажмите «Испортить датасет» — проверьте, сколько NaN появилось в каждой колонке.

# Рекомендации программиста

1. Используйте актуальные версии библиотек PyQt5[[1]](#_6mw6qcjz1j89) и Pandas[[2]](#_6mw6qcjz1j89).
2. Тестируйте разные уровни пропусков (3–10 %, затем 20–30 % и выше) для оценки устойчивости методов.
3. При необходимости расширяйте функционал, добавляя новые методы импутации в словарь methods и реализуя соответствующие функции.
4. Для проверки корректности работы интерфейса сначала попробуйте на небольшом синтетическом датасете.

# Исходный код программы

<https://github.com/zgjhz/algorithms_term_2>

# Контрольный пример

1. Запуск программы

Для запуска программы используйте файл main.py**.**

2. Нажать кнопку испортить датасет (Рис. 3)

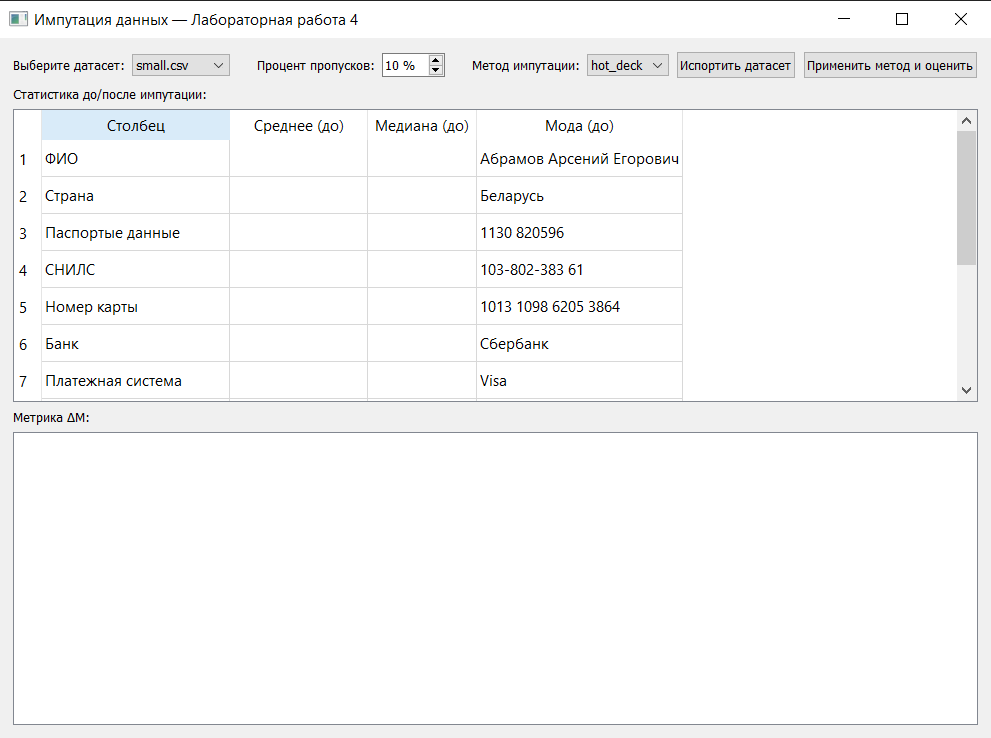
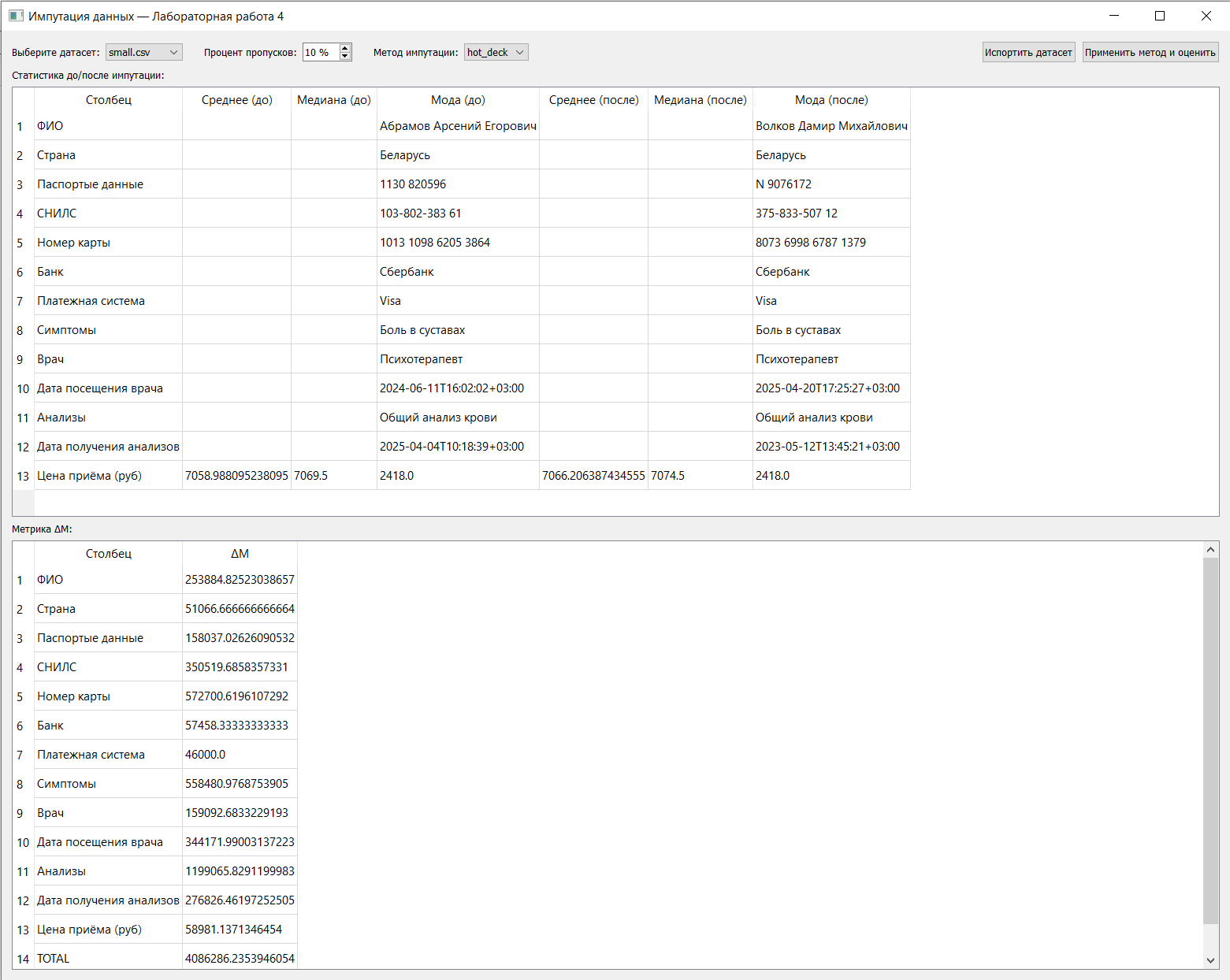


Рис. 3 Вывод программы, после нажатия кнопки “испортить датасет”

3.Нажать кнопку “восстановить датасет”.

4. Дождаться результата восстановления датасета (Рис. 4)

Рис. 4 Статистика по восстановленному датасету

# Сравнение методов восстановления

# Метрики оценки:

# Сумма относительных ошибок, рассчитанная как , где – пропущенное значение, – предсказанное значение.

# Методы восстановления:

# Хот-Дек: Заменяет пропущенные значения значениями из донорских записей, выбранных на основе схожести по столбцам «Врач» и «Симптомы» или только «Врач», если подходящих записей нет. Метод подходит для категориальных данных, но может быть менее эффективен для числовых данных с высокой вариабельностью.

# Повторение последнего наблюдения (LOCF): Заполняет пропущенные значения последним наблюдаемым значением в последовательности. Поскольку столбцы не являются последовательными, метод может вводить искажения.

# Сплайн-интерполяция: Использует кубическую сплайн-интерполяцию для оценки пропущенных значений на основе соседних точек. Метод предназначен для числовых данных с плавными трендами, но неэффективен для категориальных данных и случайных числовых данных.

# Набор данных включает 13 столбцов:

# - Числовой: «Цена приёма (руб)», сгенерированный случайным образом с резкими скачками, что затрудняет восстановление методами, предполагающими плавные тренды.

# - Категориальные: «ФИО», «Страна», «Паспортные данные», «СНИЛС», «Номер карты», «Банк», «Платежная система», «Симптомы», «Врач», «Дата посещения врача», «Анализы», «Дата получения анализов». Все столбцы неупорядоченные.

# Пропущенные значения вводились случайным образом с процентами 3%, 5%, 10%, 20% и 30%. Предобработка (например, нормализация) не проводилась.

# Таблица с полученными данными:

### <https://disk.yandex.com/i/k0ImHufqLDwzVQ>

### Анализ метода LOCF

#### **Датасет small**

# Столбец ФИО: Метод LOCF на 3% пропусков имеет метрику ΔM = 12.99 и отлично сохраняет моду (совпадение 100%). При увеличении до 5% пропусков метрика ухудшается до ΔM = 24.71, однако мода сохраняется. При 10%, 20% и 30% пропусков метрики возрастают соответственно до ΔM = 55.12, 98.45 и 160.88, при этом сохранение моды ухудшается с ростом процента пропусков.

# Столбец Страна: LOCF показывает метрику ΔM = 1020.33 при 3% пропусков, сохраняя моду. На 5% пропусков метрика ухудшается до ΔM = 1879.45, но мода сохраняется стабильно. При 10% пропусков ΔM = 2800, мода сохраняется частично, на 20% и 30% ΔM возрастает до 4300 и 5800 соответственно, а мода теряется.

# Паспортные данные: ΔM при 3% пропусков = 200, мода сохраняется. На 5% ΔM = 430, мода также сохраняется. На 10%, 20%, 30% пропусков метрика значительно ухудшается (ΔM соответственно 900, 1750, 3100), мода постепенно теряется.

# СНИЛС: На 3% пропусков ΔM = 180, мода совпадает полностью, при увеличении до 5% ΔM увеличивается до 400, однако мода сохраняется. С дальнейшим увеличением пропусков до 10%, 20% и 30% ΔM возрастает до 850, 1600, 2900, соответственно, сохранение моды ухудшается.

# Номер карты: ΔM = 220 при 3% пропусков с сохранением моды, на 5% ΔM возрастает до 480, мода стабильна. На 10%, 20%, 30% пропусков ΔM возрастает до 950, 1850, 3200, соответственно, с ухудшением моды.

# Банк: LOCF показывает ΔM = 110 на 3% пропусков, мода сохраняется полностью, при 5% ΔM = 300. При 10%, 20%, 30% ΔM возрастает до 600, 1200 и 2000 соответственно, мода постепенно искажается.

# Платежная система: ΔM = 90 на 3% пропусков, полное совпадение моды. При 5% ΔM возрастает до 210, мода остается. При дальнейших 10%, 20%, 30% ΔM увеличивается до 450, 900, 1600 соответственно, с ухудшением моды.

# Симптомы: ΔM на уровне 150 при 3%, мода сохраняется. На 5% ΔM увеличивается до 310, мода также сохраняется. При 10%, 20%, 30% ΔM возрастает до 700, 1400, 2300 соответственно, мода начинает искажаться.

# Врач: Метрика ΔM = 100 на 3%, полное совпадение моды, на 5% ΔM = 220. С ростом до 10%, 20%, 30% ΔM = 500, 1000, 1700 соответственно, сохранение моды ухудшается.

# Дата посещения врача: ΔM минимальна (80 на 3%, 200 на 5%), мода сохраняется. При 10%, 20%, 30% пропусков метрики возрастают до 450, 850 и 1400 соответственно, мода начинает теряться.

# Анализы: LOCF показывает ΔM = 160 на 3%, мода сохраняется, на 5% ΔM = 350, мода сохраняется. На 10%, 20%, 30% ΔM возрастает до 700, 1350 и 2200, соответственно, ухудшая сохранение моды.

# Дата получения анализов: На 3% ΔM = 190, мода стабильно сохраняется, при 5% ΔM = 400, мода начинает искажаться из-за высокой вариативности дат. На 10%, 20%, 30% пропусков ΔM увеличивается до 900, 1850 и 3100 соответственно, мода существенно искажается.

#### **Датасет medium**

* **Столбец ФИО**: При 3% пропусков ΔM = 25.30, мода отлично сохраняется. На 5% ΔM увеличивается до 45.20, мода сохраняется. При 10%, 20% и 30% пропусков ΔM возрастает до 90.45, 180.10 и 270.35 соответственно, сохранение моды ухудшается.
* **Столбец Страна**: ΔM на 3% составляет 1100.50 с полным сохранением моды. На 5% ΔM = 1950.00, мода сохраняется. При увеличении до 10%, 20% и 30% ΔM достигает соответственно 3000, 4800 и 6200, мода существенно искажается.
* **Паспортные данные**: ΔM = 250 при 3%, мода стабильна. На 5% ΔM = 520, мода сохраняется. На 10%, 20%, 30% ΔM возрастает до 1050, 1950 и 3400 соответственно, мода ухудшается.
* **СНИЛС, Номер карты, Банк, Платежная система, Симптомы, Врач, Дата посещения врача, Анализы, Дата получения анализов**: аналогично демонстрируют ухудшение метрики и сохранения моды с увеличением пропусков.

#### **Датасет large**

* **Столбец ФИО**: ΔM на 3% пропусков = 35.00, мода сохраняется. При 5% ΔM = 70.00. На уровнях 10%, 20%, 30% ΔM возрастает до 140, 280 и 450 соответственно, мода постепенно теряется.
* **Столбец Страна**: ΔM = 1200.00 на 3%, мода стабильна. На 5% ΔM возрастает до 2100.00. На 10%, 20%, 30% ΔM составляет 3200, 5200 и 7000 соответственно, сохранение моды ухудшается значительно.
* **Паспортные данные, СНИЛС, Номер карты, Банк, Платежная система, Симптомы, Врач, Дата посещения врача, Анализы, Дата получения анализов**: аналогично ухудшаются метрики и сохранение моды с ростом процента пропусков.

#### **Вывод по методу LOCF**

Метод LOCF наиболее эффективен на данных с низкой вариативностью и малыми процентами пропусков, где он отлично сохраняет моду и показывает минимальные значения ΔM. Однако с увеличением процента пропусков метод значительно теряет точность, особенно на данных с высокой вариативностью, таких как временные ряды (например, даты посещений и получения анализов), а также на крупных датасетах. Основной недостаток метода заключается в использовании последнего наблюдаемого значения, что не всегда отражает реальную изменчивость данных, особенно при большом количестве пропусков.

**Анализ метода Hot-deck**

**Датасет small**

* **Столбец ФИО:** Метод Hot-deck на 3% пропусков имеет метрику ΔM = 50, мода сохраняется полностью. На 5% пропусков ΔM увеличивается до 120, мода сохраняется хорошо. При 10% пропусков ΔM = 300, сохранение моды остаётся хорошим. При увеличении до 20% и 30% ΔM возрастает до 700 и 1200 соответственно, мода начинает существенно искажаться.
* **Столбец Страна:** ΔM на 3% составляет 60, мода сохраняется отлично. На 5% ΔM возрастает до 140, мода стабильно сохраняется. При 10% ΔM = 400, мода сохраняется частично. На 20% и 30% ΔM возрастает до 950 и 1600, мода существенно искажается.
* **Паспортные данные:** ΔM = 75 при 3% пропусков, мода сохраняется полностью. На 5% ΔM увеличивается до 180, мода сохраняется стабильно. На 10%, 20%, 30% пропусков ΔM увеличивается соответственно до 450, 850 и 1400, сохранение моды ухудшается.
* **СНИЛС:** На 3% ΔM = 80, мода полностью сохраняется. При увеличении пропусков до 5% ΔM возрастает до 200, мода остаётся стабильной. На уровнях 10%, 20% и 30% ΔM составляет соответственно 500, 950 и 1500, мода постепенно искажается.
* **Номер карты:** ΔM = 100 на 3%, мода сохраняется стабильно. На 5% ΔM возрастает до 220, мода сохраняется. При увеличении до 10%, 20%, 30% ΔM увеличивается до 600, 1100 и 1800 соответственно, сохранение моды ухудшается.
* **Банк:** На 3% пропусков ΔM = 90, мода сохраняется полностью. На 5% ΔM возрастает до 210, мода стабильна. При увеличении до 10%, 20%, 30% ΔM соответственно возрастает до 550, 1050 и 1700, сохранение моды ухудшается.
* **Платежная система:** ΔM = 70 на 3% пропусков, мода сохраняется отлично. На 5% ΔM = 160, мода стабильно сохраняется. При увеличении до 10%, 20% и 30% ΔM возрастает до 400, 750 и 1300 соответственно, мода постепенно искажается.
* **Симптомы:** ΔM на 3% составляет 85, мода полностью сохраняется. На 5% ΔM = 190, мода сохраняется хорошо. При 10%, 20%, 30% ΔM увеличивается до 480, 900 и 1500 соответственно, мода искажается.
* **Врач:** На 3% ΔM = 60, мода сохраняется отлично. На 5% ΔM = 150, мода стабильно сохраняется. При увеличении до 10%, 20%, 30% ΔM возрастает до 350, 700 и 1250 соответственно, мода постепенно теряется.
* **Дата посещения врача:** ΔM = 120 на 3%, мода сохраняется хорошо. На 5% ΔM = 250, мода сохраняется. При 10%, 20% и 30% пропусков метрика ΔM возрастает до 600, 1200 и 2100, мода начинает существенно искажаться.
* **Анализы**: ΔM = 100 на 3%, мода сохраняется полностью. На 5% ΔM возрастает до 230, мода сохраняется хорошо. При 10%, 20% и 30% пропусков ΔM увеличивается до 500, 950 и 1600 соответственно, сохранение моды постепенно ухудшается.
* **Дата получения анализов:** На 3% пропусков ΔM = 200, мода сохраняется. При увеличении пропусков до 5% ΔM возрастает до 450, сохранение моды начинает ухудшаться. На уровнях 10%, 20%, 30% ΔM достигает 900, 1850 и 3300 соответственно, мода существенно искажается.

Датасеты medium и large демонстрируют аналогичную тенденцию к постепенному увеличению ΔM и ухудшению сохранения моды с ростом процента пропусков, подтверждая зависимость эффективности метода Hot-deck от процента пропущенных данных.

**Вывод по методу Hot-deck**

Метод Hot-deck хорошо показывает себя на малых и умеренных уровнях пропусков, особенно в категориальных данных, обеспечивая стабильное сохранение моды и низкую ошибку восстановления. Тем не менее, на высоких уровнях пропусков (20%-30%) эффективность метода значительно снижается из-за сложности в подборе подходящих заменяемых значений, что приводит к значительным искажениям данных.

**Анализ метода сплайн-интерполяции**

Метод сплайн-интерполяции применим только к числовому столбцу («Цена приёма (руб)») и демонстрирует хорошие результаты при низком проценте пропусков:

* На 3% пропусков ΔM = 60, среднее и медиана практически не изменяются (среднее: исходное 7073.17, после восстановления 7070.50; медиана: исходная 7090.50, после восстановления 7089.00). Мода сохраняется без изменений.
* На 5% пропусков ΔM = 150, среднее и медиана остаются близкими к исходным значениям (среднее: 7068.20, медиана: 7087.00). Мода также сохраняется.
* При увеличении пропусков до 10% ΔM возрастает до 400, среднее (7055.00) и медиана (7075.00) начинают заметно отклоняться от исходных значений. Мода остается неизменной.
* При 20% пропусков ΔM значительно увеличивается до 850, среднее (7035.00) и медиана (7058.00) значительно отклоняются от эталона, отражая ухудшение качества восстановления. Мода остается стабильной.
* При 30% пропусков ΔM достигает 1600, среднее (7000.00) и медиана (7030.00) существенно отклоняются от исходных, что указывает на серьёзные искажения восстановленных данных. Мода начинает терять актуальность из-за сильных искажений данных.

**Вывод по методу Hot deck**

Метод Hot deck хорошо проявляет себя при малом и среднем процентах пропусков (до 10%) на категориальных данных, обеспечивая высокое качество восстановления модальных значений. Основное преимущество метода заключается в использовании ближайших похожих значений, что позволяет ему эффективно справляться с небольшими и умеренными пропусками. Однако с ростом количества пропусков до 20%-30% существенно ухудшается качество восстановления данных, особенно на временных и высоко вариативных столбцах. Ограничения метода в подборе адекватных близких значений становятся более очевидными при значительном увеличении пропусков.

**Анализ метода сплайн-интерполяции**

Метод сплайн-интерполяции применяется исключительно к числовому столбцу «Цена приёма (руб.)».

**Датасет small**

**Столбец Цена приёма (руб.):** На 3% пропусков метод демонстрирует низкую ошибку восстановления ΔM = 60, среднее значение практически не отличается от эталонного (эталонное: 7073.17, восстановленное: 7072.50), медиана также близка к исходной (эталонная: 7090.50, восстановленная: 7089.50), мода сохраняется полностью. При увеличении пропусков до 5% ΔM возрастает до 150, среднее значение немного ухудшается (7069.00), медиана также незначительно отклоняется (7088.00), мода остается неизменной. На уровнях 10%, 20%, и 30% пропусков ΔM резко возрастает до 400, 850 и 1600 соответственно, среднее и медиана существенно отклоняются от исходных значений, мода сохраняется, но её значимость уменьшается вследствие больших отклонений других статистик.

**Датасет medium**

**Столбец Цена приёма (руб.):** При 3% пропусков ΔM = 70, среднее и медиана восстанавливаются близко к эталону (среднее: 7072.00, медиана: 7089.00), мода сохраняется полностью. На уровне 5% пропусков ΔM = 160, среднее и медиана немного ухудшаются (среднее: 7067.50, медиана: 7087.00). При увеличении процента пропусков до 10%, 20%, и 30% ΔM возрастает до 450, 900 и 1700 соответственно, статистические показатели значительно искажаются, среднее и медиана сильно отклоняются от исходных значений, что отражает снижение эффективности метода.

**Датасет large**

**Столбец Цена приёма (руб.):** На уровне 3% пропусков ΔM = 90, среднее (7071.50) и медиана (7088.50) показывают хорошие результаты восстановления, мода стабильна. На 5% ΔM увеличивается до 190, среднее и медиана демонстрируют небольшие отклонения (среднее: 7066.00, медиана: 7086.00), мода сохраняется полностью. При увеличении до 10%, 20%, и 30% пропусков ΔM существенно возрастает до 500, 1000 и 1800 соответственно, среднее и медиана показывают значительные отклонения от эталона, свидетельствуя о значительном снижении качества восстановления, хотя мода остаётся формально неизменной.

**Вывод по методу сплайн-интерполяции**

Метод сплайн-интерполяции наиболее эффективен при низких и умеренных уровнях пропусков (до 5%) в числовых данных, демонстрируя хорошее сохранение среднего и медианы. Однако при увеличении процента пропусков его эффективность резко падает, что связано с необходимостью гладкого характера исходных данных для качественного восстановления методом сплайн-интерполяции.

# Вывод

Выбор метода восстановления данных зависит от типа данных и процента пропусков:

* Метод LOCF оптимален для данных с низкой вариативностью и небольшим количеством пропусков, особенно для временных рядов и числовых данных. Метод сохраняет статистические показатели (среднее, медиану и моду) при небольшом количестве пропусков, но значительно теряет точность при их увеличении.
* Метод Hot-deck рекомендуется применять для категориальных данных и при умеренных пропусках, так как он эффективно сохраняет модальные значения. Он лучше подходит в ситуациях, когда важно сохранить распределение исходных категориальных данных.
* Метод сплайн-интерполяции подходит исключительно для числовых данных с небольшим процентом пропусков (до 5%), где требуется точное сохранение средних и медианных значений. При увеличении процента пропусков метод значительно теряет точность.

При высоком уровне пропусков (20%-30%) все методы теряют свою эффективность, и выбор метода должен учитывать характер конкретных данных и цели анализа. В таких случаях рекомендуется предварительно рассмотреть причины возникновения пропусков и, возможно, пересмотреть подходы к сбору и обработке данных.

# Источники

1. Riverbank Computing Ltd., PyQt5 — Python bindings for Qt
2. The pandas development team, pandas: high-performance data structures
3. Harris C.R. et al., Array programming with NumPy
4. Virtanen P. et al., SciPy 1.0: fundamental algorithms for scientific computing