Базовые понятия статистики

Работа с пропусками и выбросами

Олег Булыгин

Python для анализа данных

Олег Булыгин

IT-аудитор, ментор и наставник

Аккаунты в соц.сетях

fb.com/obulygin91

linkedin.com/in/obulygin

vk.com/obulygin91

@obulygin91

obulygin91@ya.ru

Содержание

1 Статистика и ее задачи

2 Типы данных

3 Основные понятия статистики

4 Выбросы

5 Пропущенные значения

Статистика – это … ?

Статистика

– это отрасль знаний, которая занимается общими вопросами сбора, изменения, мониторинга и анализа данных

Два направления статистического анализа данных

1

Описательная статистика

description statistics

занимается обработкой данных, их систематизацией, наглядным представлением в форме графиков и таблиц, а также их количественным описанием посредством основных статистических показателей.

2

Индуктивная статистика

inferential statistics

занимается обобщением информации о выборке для получения представлений о свойствах генеральной совокупности

Генеральная совокупность и выборка

Генеральная совокупность

это совокупность всех объектов или наблюдений, относительно которых исследователь намерен делать выводы при решении конкретной задачи.

Выборка

часть генеральной совокупности элементов, которая охватывается задачами нашего анализа.

Репрезентативная выборка

выборка, характеристики который соответствуют характеристикам генеральной совокупности.
Репрезентативность позволяет переносить выводы о выборке на выводы о всей совокупности.

Задача из какой сферы?

Имеются данные о 100 000 просмотрах рекламного баннера. Необходимо определить, какой процент людей кликнул на него

Описательная статистика	Индуктивная статистика

Задача из какой сферы?

Имеются данные о 100 000 просмотрах рекламного баннера. Необходимо определить, какой процент людей кликнул на него

Описательная статистика

Индуктивная статистика

Задача из какой сферы?

После опроса 100 клиентов магазина нужно определить, какой процент всех клиентов довольны продуктов компании

Описательная статистика

Индуктивная статистика

Задача из какой сферы?

После опроса 100 клиентов магазина нужно определить, какой процент всех клиентов довольны продуктов компании

Описательная статистика

Индуктивная статистика

Задача из какой сферы?

Мы поймали в реке 20 рыб. Какой средний вес рыб во всей реке?

Описательная статистика	Индуктивная статистика

Задача из какой сферы?

Мы поймали в реке 20 рыб. Какой средний вес рыб во всей реке?

Описательная статистика

Индуктивная статистика

Задача из какой сферы?

Есть данные обо всех обращениях клиентов в техническую поддержку. Необходимо определить среднее время ответа на их обращения

Описательная статистика

Индуктивная статистика

Задача из какой сферы?

Есть данные обо всех обращениях клиентов в техническую поддержку. Необходимо определить среднее время ответа на их обращения

Описательная статистика

Индуктивная статистика

Типы данных

КОЛИЧЕСТВЕННЫЕ

ДИСКРЕТНЫЕ
(DISCRETE)

выражены ограниченным набором значений (“их можно пронумеровать”)

НЕПРЕРЫВНЫЕ
(CONTINUES)

выражены неограниченным набором значений (“не имеют конечной точности измерения”)

КАЧЕСТВЕННЫЕ

ПОРЯДКОВЫЕ
(ORDINAL)

ранжирование значимо

НОМИНАЛЬНЫЕ
(NOMINAL)

ранжирование незначимо (сравнение не имеет смысла)

Какая это величина?

Киловатт-час электроэнергии

Дискретная величина	Непрерывная величина	Категориальная величина

Какая это величина?

Киловатт-час электроэнергии

Дискретная величина	Непрерывная величина	Категориальная величина

Какая это величина?

Температура воздуха

Дискретная величина	Непрерывная величина	Категориальная величина

Какая это величина?

Температура воздуха

Дискретная величина	Непрерывная величина	Категориальная величина

Какая это величина?

Количество единиц товара на складе

Дискретная величина	Непрерывная величина	Категориальная величина

Какая это величина?

Количество единиц товара на складе

Дискретная величина	Непрерывная величина	Категориальная величина

Какая это величина?

Почтовый индекс

Дискретная величина	Непрерывная величина	Категориальная величина

Какая это величина?

Почтовый индекс

Дискретная величина	Непрерывная величина	Категориальная величина

Какая это величина?

Количество пройденных курсов в Нетологии

Дискретная величина	Непрерывная величина	Категориальная величина

Какая это величина?

Количество пройденных курсов в Нетологии

Дискретная величина	Непрерывная величина	Категориальная величина

Какая это величина?

Марка автомобиля

Дискретная величина	Непрерывная величина	Категориальная величина

Какая это величина?

Марка автомобиля

Дискретная величина	Непрерывная величина	Категориальная величина

Какая это величина?

Количество кликов по рекламному баннеру

Дискретная величина	Непрерывная величина	Категориальная величина

Какая это величина?

Количество кликов по рекламному баннеру

Дискретная величина	Непрерывная величина	Категориальная величина

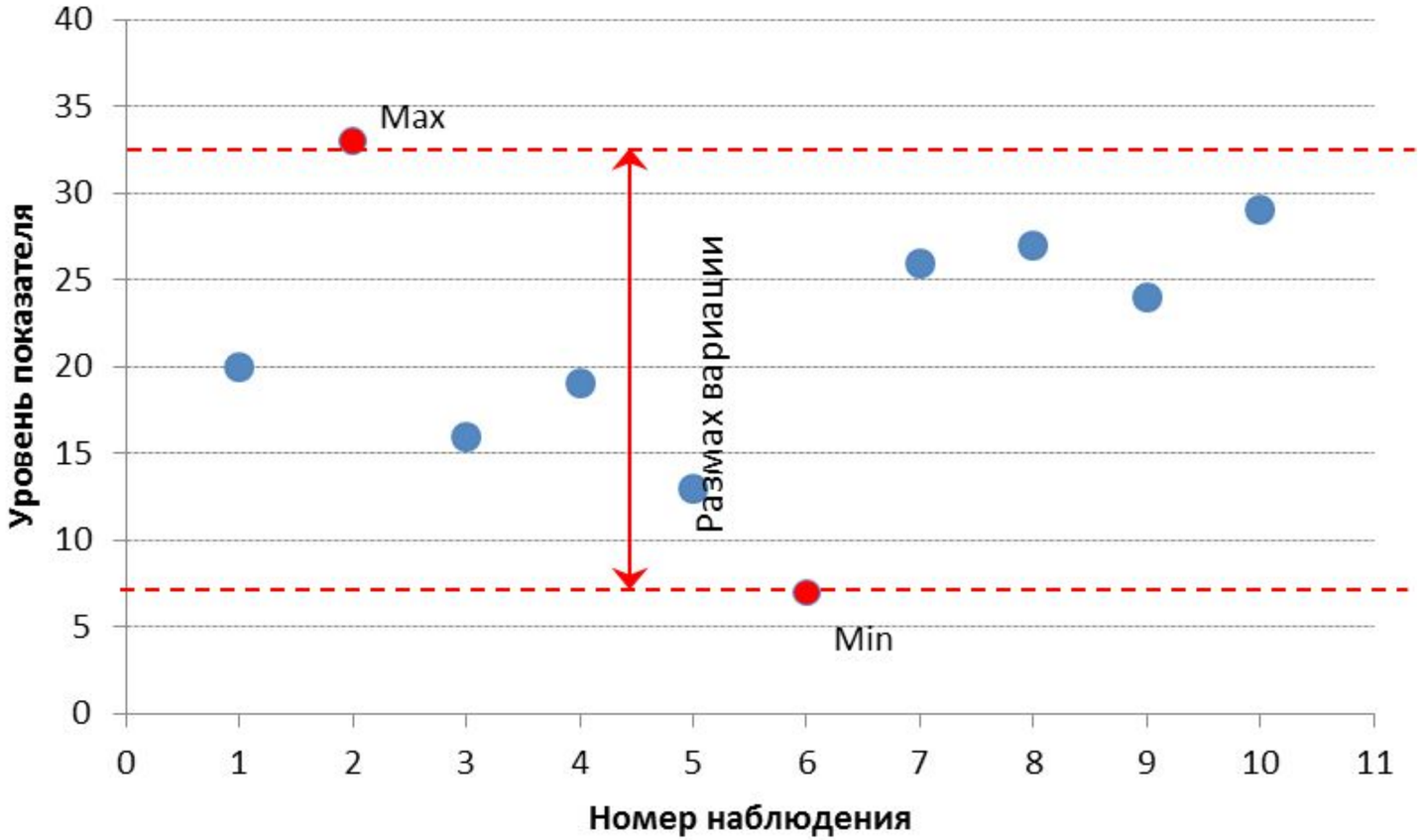
Очевидные штуки :)

Максимальная величина

Минимальная величина

Размах

разность между наибольшим и наименьшим значениями показателя



Меры центральной тенденции

числа, характеризующие выборку по уровню выраженности величины

Среднее арифметическое

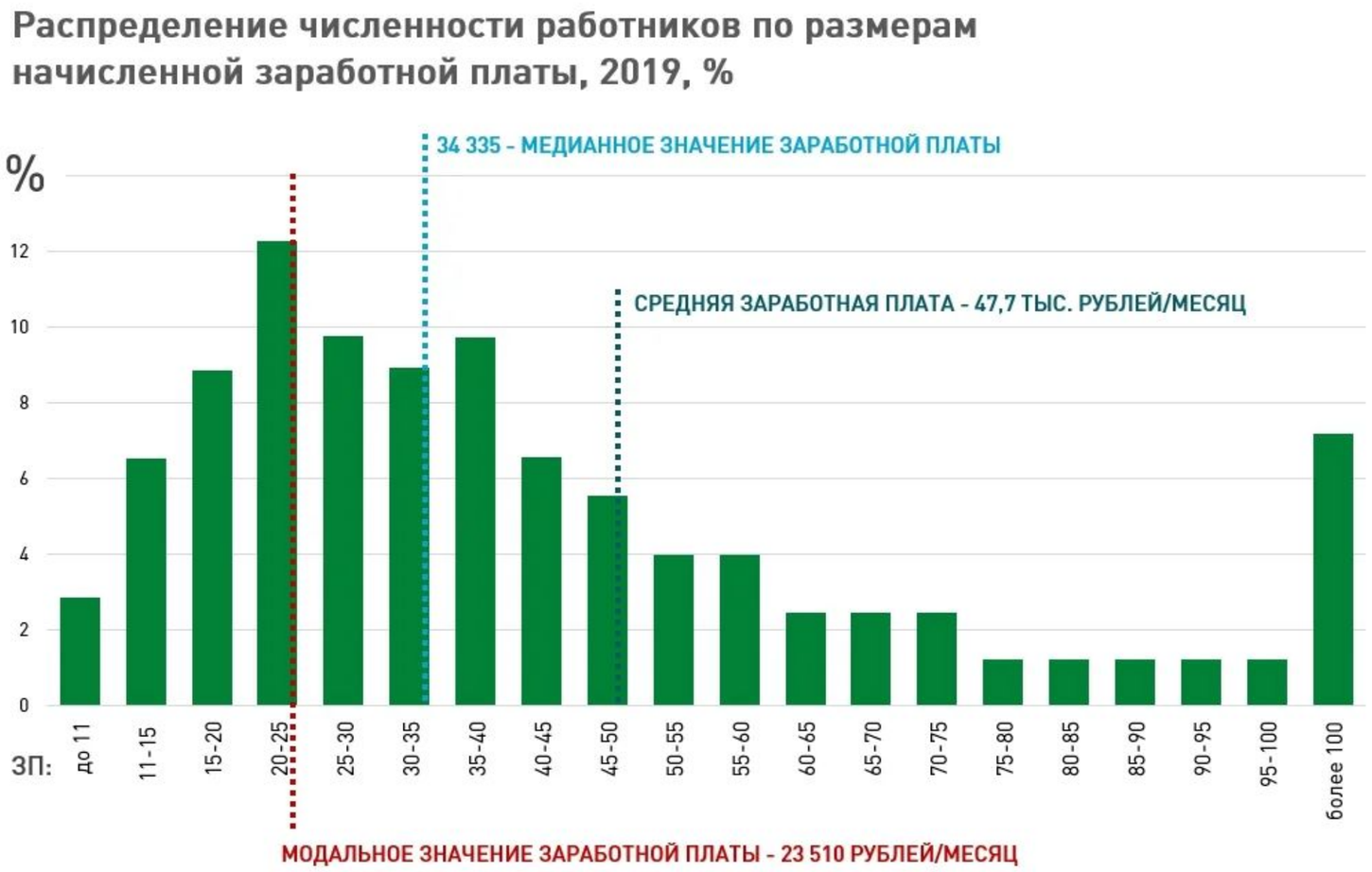
частное от деления суммы всех чисел ряда на их количество

Мода

число, наиболее часто встречающееся в данном ряду

Медиана

число, половина из элементов выборки больше которого, а другая половина – меньше



Меры разброса

характеризуют степень индивидуальных отклонений величины от среднего

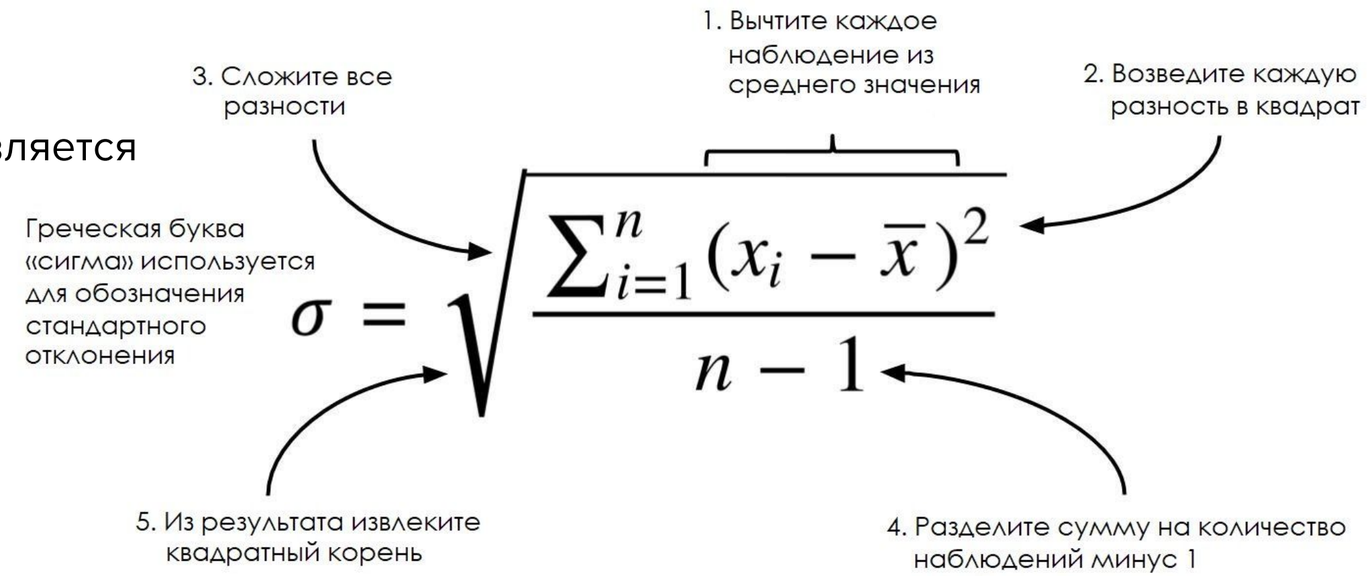
Стандартное отклонение

среднее квадратическое отклонение, среднеквадратичное отклонение, квадратичное отклонение

измеряется в единицах самой случайной величины и используется при расчете стандартной ошибки среднего арифметического, при построении доверительных интервалов, при статистической проверке гипотез, при измерении линейной взаимосвязи между случайными величинами. Является корнем из дисперсии.

Дисперсия

просто квадрат стандартного отклонения. Во многих статистических формулах удобнее использовать СКО, а не извлекать каждый раз корень из дисперсии.



Меры разброса

характеризуют степень индивидуальных отклонений величины от среднего

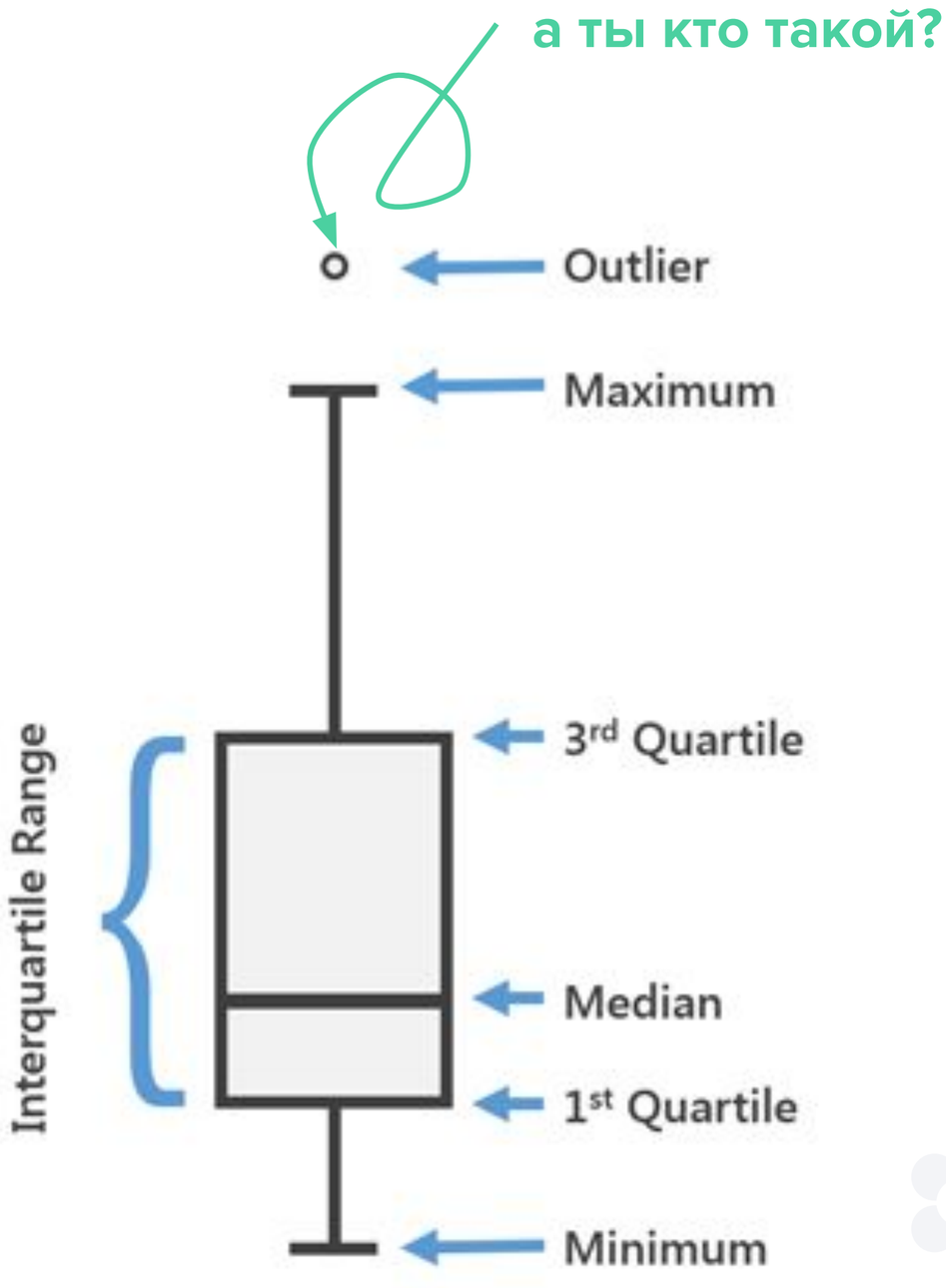
Квантили (процентили)

это показатель, показывающий значение, ниже (выше) которого падает определенный процент наблюдений в группе наблюдений.

- 0.25-квантиль называется первой (или нижней) квартилью;
- 0.5-квантиль называется второй квартилью (это же тоже самое, что медиана!);
- 0.75-квантиль называется третьей (или верхней) квартилью.

Межквартильный размах (IQR)

это разница между 1-м и 3-м квартилями, т.е. между 25-м и 75-м процентилями.



Выбросы (outliers)

результаты измерения, сильно выделяющиеся в общей выборке

Наиболее простой методов обнаружения выбросов основан на межквартильном размахе (т.е. все что не попадает в указанные диапазоны, является выбросом):

$$\text{Lower Outlier} = Q1 - (1.5 \times IQR)$$

$$\text{Higher Outlier} = Q3 + (1.5 \times IQR)$$

Наиболее распространенные причины выбросов в наборе данных:

- Ошибки ввода данных (**человеческий фактор**)
- **Погрешности измерения** (ошибки приборов)
- **Экспериментальные ошибки**
- **Преднамеренное** (например, сделанные для проверки методов обнаружения оборудованием)
- **Ошибки обработки данных**
- **Ошибки выборки** (извлечение или смешивание данных из неправильных или различных источников)
- **Естественные выбросы** (не ошибки, а реальные исключительные наблюдения в данных

Влияние выбросов на данные

1. приводят к различным проблемам во время статистического анализа;

2. могут оказывать существенное влияние на базовые статистики, характеризующие выборку (например, среднее и стандартное отклонение).

Существует ряд статистик, таких как медиана, которые можно считать робастными (устойчивыми) к наличию выбросов. Поэтому всегда стоит выбирать статистику, которая лучше описывает данные (например, среднее арифметическое под влиянием выбросов может сильно искажать представление о данных).

Выбор способа работы с выбросами существенно зависит от специфики набора данных и целей проекта. В целом их обработка во многом похожа на обработку пропущенных значений. Можно удалить записи или признаки с выбросами, либо скорректировать их, либо оставить без изменений.

Пропущенные значения

Почему они бывают? Причин может быть много:

- данных просто нет (мы их не знаем);
- отсутствие данных имеют естественную причину и объяснимо;
- человеческая ошибка сбора/ввода;
- технические ошибки и проблемы, которые привели к потере данных.

Реальные данные почти всегда содержат пропуски.
Нужно решить, что с ними делать исходя из причины их возникновения.
Неудачный выбор метода заполнения пропусков может не только не улучшить, но и сильно ухудшить результаты анализа.

Пропущенные значения

Для того чтобы понять, как правильно обработать пропуски, необходимо определить **механизмы их формирования**.

MCAR

Missing Completely At Random

механизм формирования пропусков, при котором вероятность пропуска для каждого наблюдения одинакова. В таком случае игнорирование/исключение записей, содержащих пропущенные данные, не ведет к искажению результатов. Замена допустима.

MAR

Missing At Random

на практике данные обычно пропущены не случайно, а ввиду некоторых закономерностей (если вероятность пропуска может быть определена на основе другой имеющейся в наборе данных информации, не содержащей пропуски). В таком случае удаление или правильная замена пропусков также не приведет к *существенному* искажению результатов.

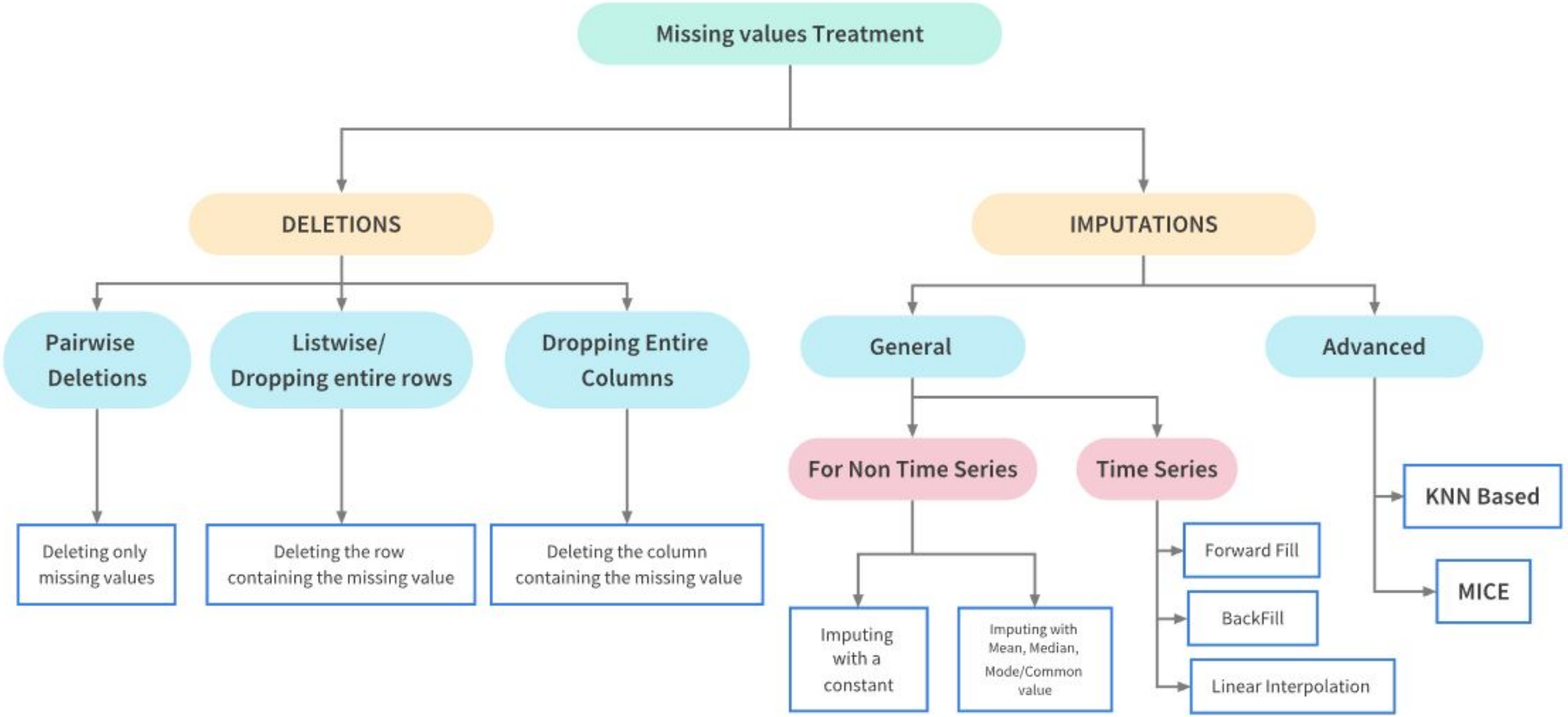
MNAR

Missing Not At Random

механизм формирования пропусков, при котором данные отсутствуют в зависимости от неизвестных факторов. MNAR предполагает, что вероятность пропуска могла бы быть описана на основе других атрибутов, но информация по этим атрибутам в наборе данных отсутствует.
В таком случае любые манипуляции с пропусками могут привести к *существенному* искажении выводов.

Важно! В учебных целях мы пока будем предполагать, что наши пропуски всегда полностью случайны.

Простейшие методы работы с пропусками



Игнорирование/удаление пропусков

1. **Все операции в pandas по-умолчанию просто игнорируют пропуски.**
2. Удаление строк, содержащих пропуски при MCAR не приведет к существенному искажению свойств данных. Но при MAR и, особенно, при MNAR смещение статистических свойств выборки могут быть значительными. В случаях, когда пропусков в данных много, это становится ощутимой проблемой и может сильно исказить результаты анализа.

Удаление столбцов с пропусками подходит только в том случае, если недостающие данные не являются информативными и пропусков чрезвычайно много (> 80%).

Заполнение пропусков

Заполнение константой

Замена пропущенных значений константой, которая заведомо не может попадать в реальные значения (-999, -1, “Нет информации” и пр.) позволит сгруппировать пропуски и рассматривать их как отдельную категорию.

Может быть полезно, когда не нужно делать никаких агрегированных расчетов и прогнозов, а наблюдения с пропусками есть.

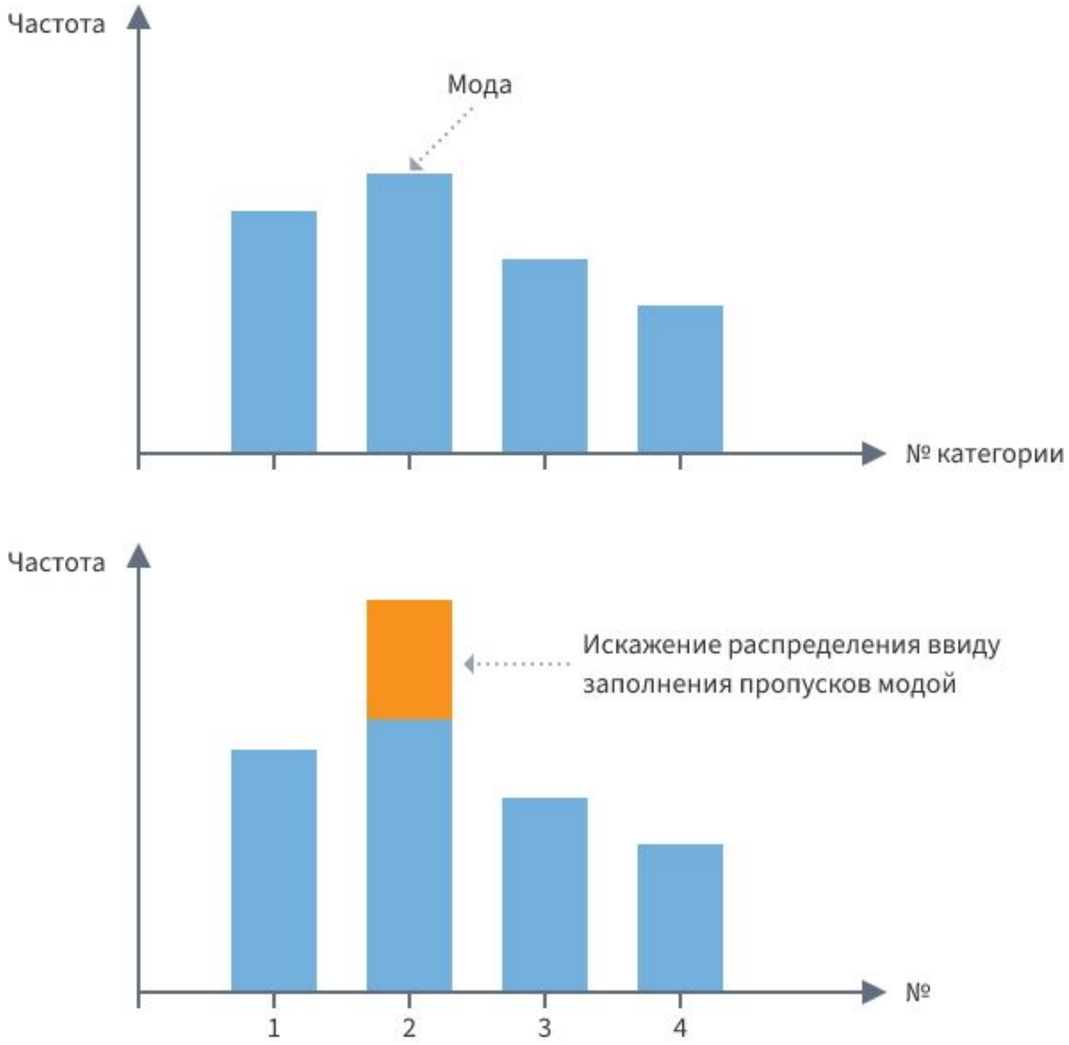
Заполнение пропусков (2)

Заполнение средним/модой/медианой

Для числовых признаков можно заполнить пропуски средним или медианным значением, полученное из остальных записей. В случае категориальной дискретной характеристики наиболее часто используется заполнение модой.

Это позволит нам не терять данные, но при этом может внести определенные искажения в выводы.

Лучше разумно подходить к заполнению пропусков и заполнять не просто средними, а средними по какой-то группе.



Заполнение пропусков (3)

Заполнение следующим/ предыдущим значением

Данный метод применяется, как правило, при заполнении пропусков во временных рядах, когда последующие/предыдущие значения априори сильно взаимосвязаны с предыдущими.

При этом данный метод тоже может привести к существенным искажениям статистических свойств даже в случае MCAR.

Возможна ситуация когда применение приведет к дублированию выбросов (заполнению пропусков аномальным значением). А если в данных много последовательно пропущенных значений, то использование метода также приводит к неправильным результатам.

Базовые понятия статистики

Вопросы?

Олег Булыгин

Python для анализа данных

нетология