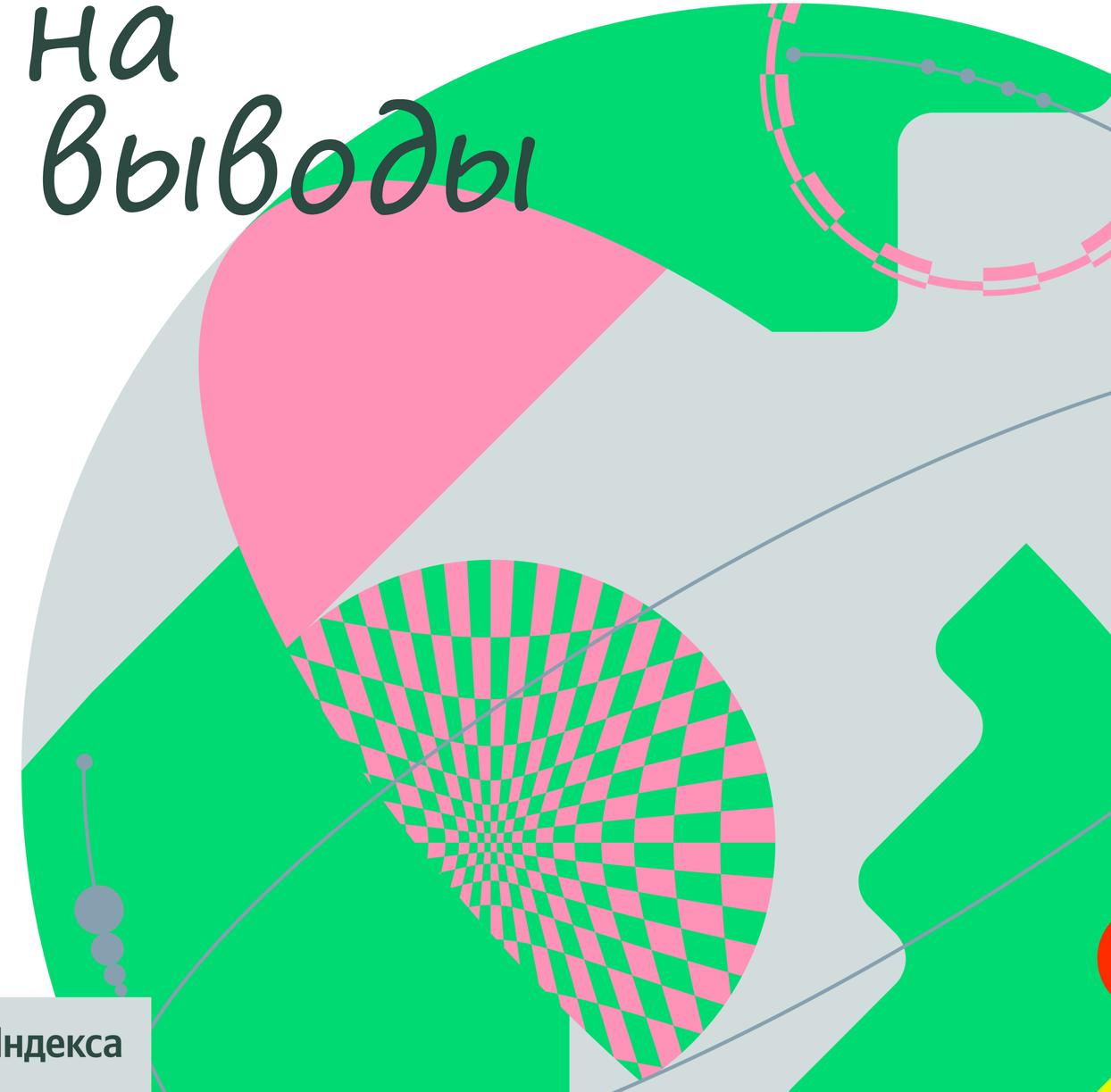
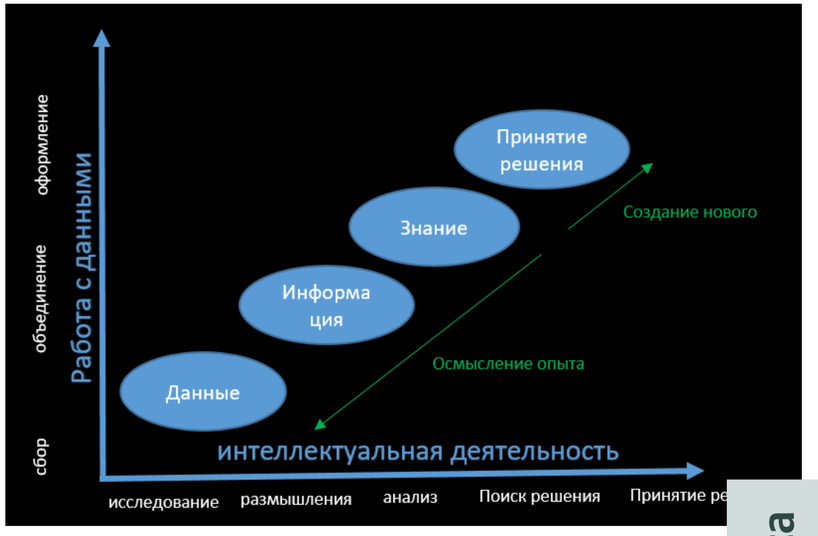
Как качество исходных данных влияет на аналитические выводы



Академия Яндекса

Ребята, эта глава – чуть ли не самая важная!!!!
Помните, на первой встрече мы рассматривали вот эту картинку:



И тут в явном виде видно, что данные — это фундамент, на котором держится компания, если она опирается на аналитическую культуру. А если менеджеры, принимающие решения, не располагают своевременной, релевантной и достоверной информацией, у них не останется другого выхода, как только положиться на собственную интуицию, что не всегда будет лучшее решение. Поэтому, качество данных — ключевой момент и в аналитике и при построении прогнозных моделей.

Если вы аналитик, то вам нужны правильные данные, собранные правильным образом и в правильной форме, в правильном месте, в правильное время.

Если какое-то из этих требований не выполнено или выполнено недостаточно хорошо, то у вас, как правило, снижается качество выводов, которые сможете сделать на основании данных. А в целом – уменьшается круг вопросов, на которые возможно дать ответ.

Вот, о чём будем говорить далее:

1. **Концептуальные моменты качества данных. Как бы излишне философски не выглядела эта тема, но она нужна, т.к. осмысление её помогает обеспечить правильность процесса сбора данных.**

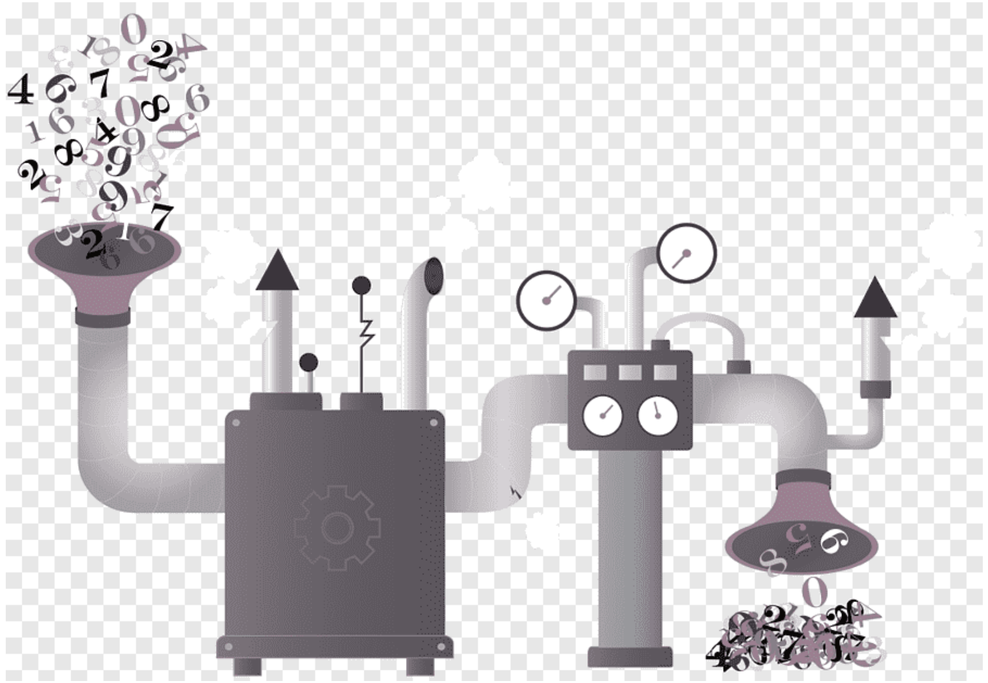
2. **Обсудим жизненный путь данных. Это позволит понять на каких этапах и процессах стоит заострять внимание**

3. **Рассмотрим практические инструменты для анализа качества данных.**

Garbage In, Garbage Out

Какие данные введешь, такие и получишь

Мусор на входе, мусор на выходе



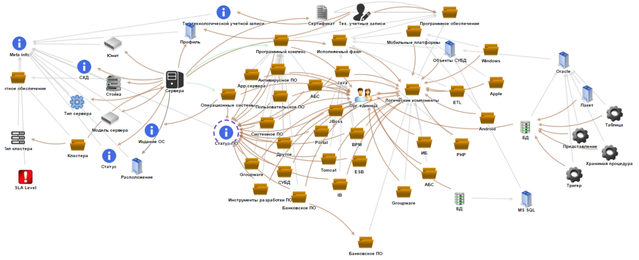
Концептуальные моменты качества данных

Доступность

У аналитика должен быть доступ к данным. Чаще всего, это предполагает разрешение на их получение (доступы) и умение их доставать, т.е. владение инструментом SQL. Сейчас у компаний есть понимание этого момента. Поэтому, одним из требований при приёме на работу аналитиком, является знание SQL.

Точность

Данные должны отражать истинные значения или положение дел. Например, показания неправильно настроенного датчика температуры, ошибка в дате рождения или устаревший адрес — это все примеры неточных данных. На производстве есть отдельная статистико-аналитическое направление MSA (**Анализ измерительных систем**). Существуют ГОСТы на этот счёт.



Взаимосвязанность

Должна быть возможность точно связать одни данные с другими. Например, заказ клиента в интернет-магазине должен быть связан с информацией о нем самом, с товаром или товарами из заказа, с платежной информацией и информацией об адресе доставки и т.д. Этот набор данных обеспечивает полную картину заказа клиента. Тут понятно.

Академия Яндекса

Концептуальные моменты качества данных

Полнота

Под неполными данными может подразумеваться как отсутствие части информации (например, в сведениях о клиенте не указано его имя), так и полное отсутствие единицы информации (например, в результате ошибки при сохранении в базу данных потерялась вся информация о клиенте)

Непротиворечивость

Данные должны быть согласованными. Например, адрес конкретного клиента в одной базе данных должен совпадать с адресом этого же клиента в другой базе.

При наличии разногласий один из источников следует считать основным или вообще не использовать сомнительные данные до устранения причины разногласий. Однозначность Каждое поле, содержащее индивидуальные данные, имеет определенное, недвусмысленное значение. Четко названные поля в совокупности со словарем базы помогают обеспечить качество. Отчасти PEP8 об этом. Но, всё же, аналитик сам должен следить за этим.

Концептуальные моменты качества данных

Релевантность

Данные зависят от характера анализа. Например, данные о продажах Сытного рынка в Петербурге 1900 года может быть познавательно, но никак не относиться к текущим продажам современной интернет-торговли.

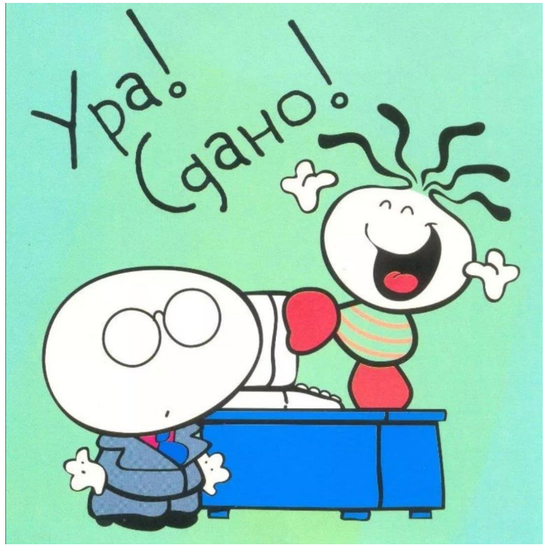
Надёжность

Данные должны быть одновременно полными (то есть содержать все сведения, которые вы ожидали получить) и точными (то есть отражать достоверную информацию).

Своевременность

Между сбором данных и их доступностью для использования в аналитической работе всегда проходит время. На практике это означает, что аналитики должны получать данные как раз вовремя, чтобы завершить анализ к необходимому сроку. А если сроки отдачи данных из базы будут 4-5 недель, то … может быть поздно.

Как видите, хоть и много слов произнесли, что не соблюдение или ошибки в одном из этих аспектов может привести к тому, что данные окажутся частично или полностью непригодными к использованию или, что хуже, будут казаться достоверными, но приведут к неправильным управленческим решениям.



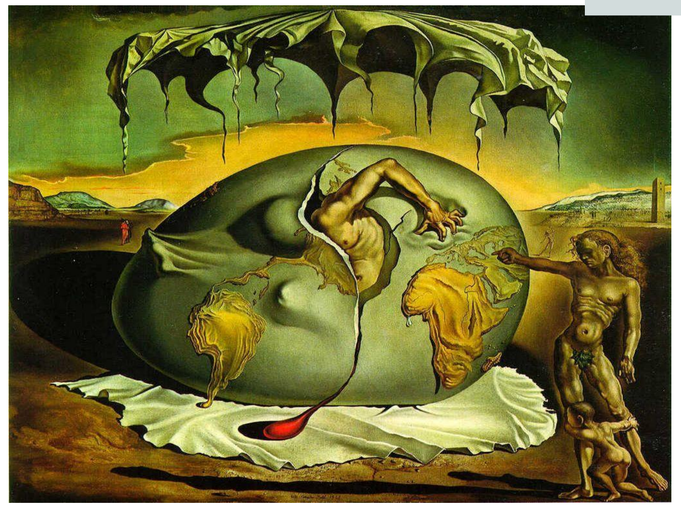
На каком этапе могут появляться ошибки

1. Генерация данных

Первый шаг в рождении данных. Поэтому, начнём с него.
Появляется в результате факторов:
- технологического (приборы)
- программного (сбои)
- человеческого.

В случае технологического фактора, приборы могут быть настроены неправильно, что может сказаться на полученных данных. Например, датчик температуры показывает 48 °C вместо 44 °C на самом деле.

Как говорилось выше, есть специальное направление «Анализ измерительной системы». Оно включает в себя:
- сертификация оборудования
- проверка методологии измерения
- качество операторов-измерителей.



На каком этапе могут появляться ошибки

2. Ввод данных

Когда данные генерируются вручную, например, при работе клиник – измерение роста и веса пациента, их необходимо зафиксировать. Большой объем данных сегодня по-прежнему сначала попадает на бумагу в качестве промежуточного шага до попадания в компьютер. И вот на этом этапе может возникнуть множество ошибок.

В целом ошибки при вводе информации можно свести к четырем типам.

Запись: введенные слова или показатели не те, что были в оригинале.
Вставка: появление дополнительного символа: 56,789 → 564,789.
Удаление: один или несколько символов теряются: 56,789 → 56,89.
Перемена мест: два или более символов меняются местами: 56,789 → 56,798.

(В качестве отдельных категорий «Вставки» и «Удаления» можно выделить:
○ диттографию — случайное повторение символа (56,789 → 56,7789)
○ гаплографию — пропуск повторяющегося символа (56,779 → 56,79).
Эти термины употребляют ученые, занимающиеся восстановлением поврежденных и переписанных от руки древних текстов, и обозначают разновидность проблемы с некачественными данными.)

На каком этапе могут появляться ошибки

Как бороться с этим:

1) Первый шаг заключается в сокращении количества этапов от генерации данных до ввода: если есть возможность избежать бумажной формы, лучше сразу вносить данные в компьютер. Везде, где возможно, стоит добавлять проверку значения каждого поля в свою электронную форму

2) Нужно стремиться к тому, чтобы пользователю пришлось вводить как можно меньше данных: лучше предложить варианты ответа на выбор, если, конечно, это позволяет формат требуемой информации.

3) Если есть свободные руки – параллельная проверка

4) При передаче важных данных в цифровой форме, например номеров банковских счетов, номеров социальной страховки и т.д. используют метод «контрольное число». После передаваемого номера добавляется число, которое представляет собой определенную функцию остальных цифр номера, и это число используется для проверки того, что предыдущие цифры были переданы из системы в систему без ошибок. Предположим, вам нужно передать индекс 12149. Воспользуемся самой простой схемой. Последовательно сложим все цифры, составляющие наш индекс, и получим 17. Сложим и эти цифры, получим 8. (хотя это метод не исключает ошибок перестановок)

В общем, есть методы, минимизирующие ошибки генерации и ввода данных. Но всё же, 100% защиты, пока, нет.

Следующий фильтр по данным, может быть на этапе разведочного анализа данных. 

Разведочный анализ данных

1.Описательный анализ количественных и качественных данных.

(Мин/макс, Квартили, средняя/медиана, графики – гистограмма и ящик с усами . Мониторим вопрос: адекватны ли данные?)

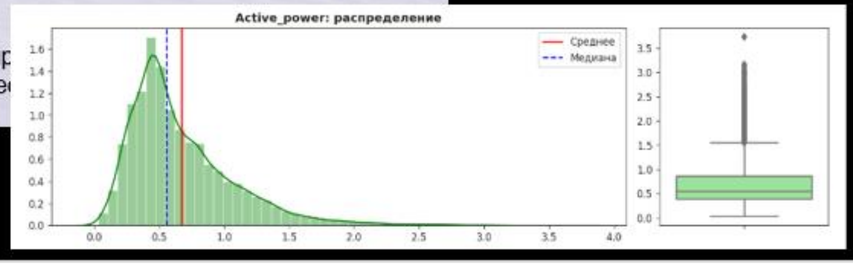
Зло ли выбросы?

Статистический ответ

+-1.5 *H подозрения на выброс +-3*H супер выброс

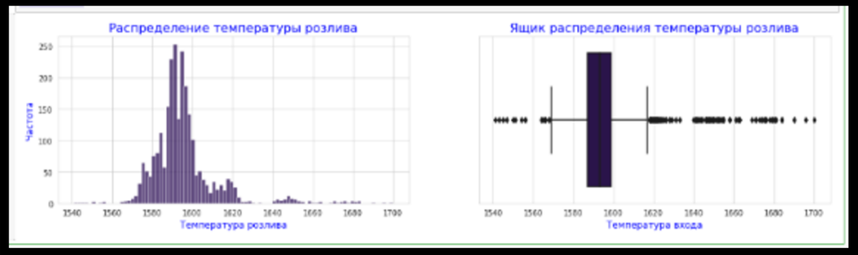
ВЫБРОС:
1) ОШИБКА ВВОДА
2) ЧТО-ТО ОСОБЕННОЕ В ДАННЫХ

Используется:
•для оценки нормальности распр
•для сравнительного анализа не



Зло ли выбросы?

Ответ жизни



В промышленности: выброс – не выброс, а реальная производственная ситуация. Просто встречается редко.

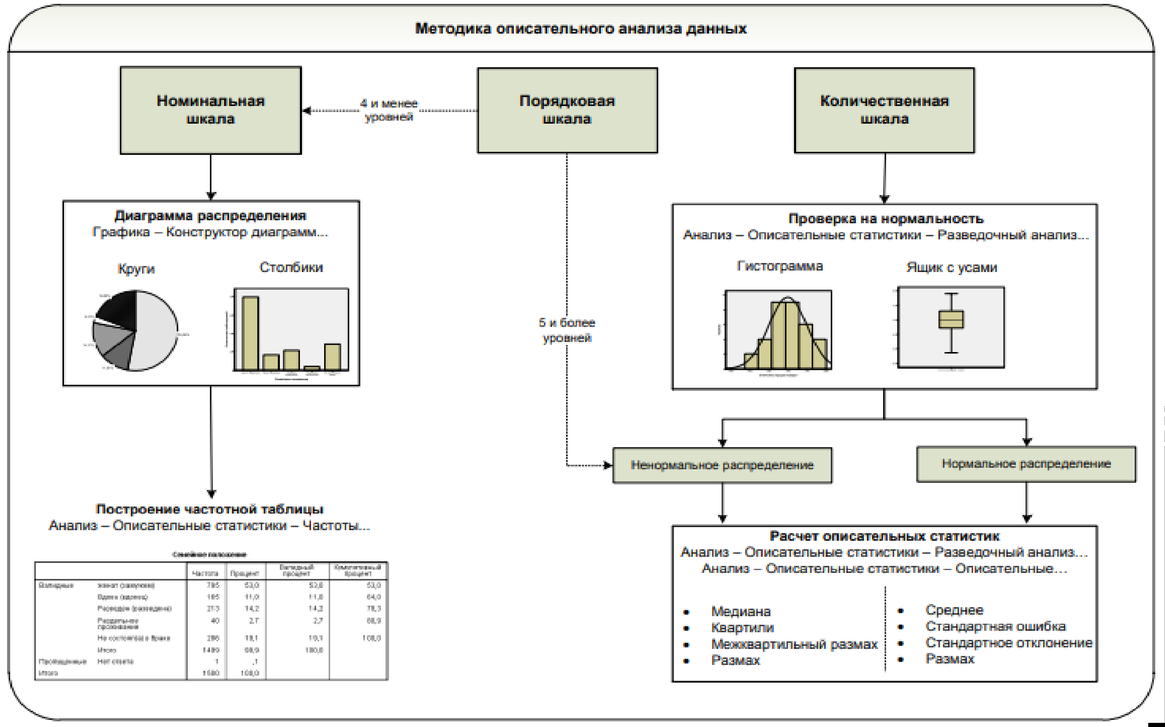
В целом: есть облако данных, которые являются просто самостоятельным сегментом/группой для исследования.

В медицине: вообще исключают выбросы как понятие: есть реальная ситуация с пациентом

Разведочный анализ данных

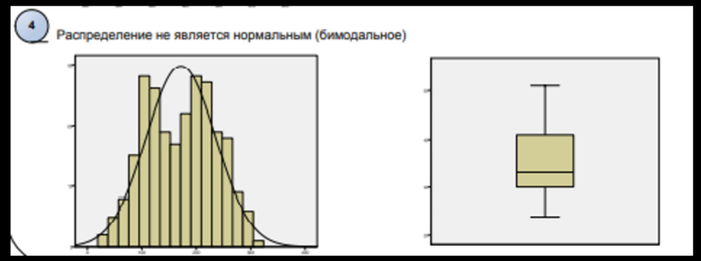
1.Описательный анализ количественных и качественных данных.

(Мин/макс, Квартили, средняя/медиана, графики – гистограмма и ящик с усами . Мониторим вопрос: адекватны ли данные?)



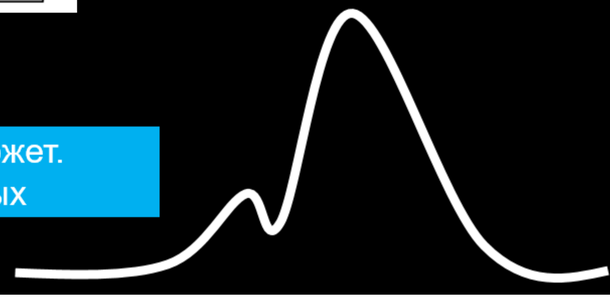
ЗАЧЕМ (примеры)

- Нормальное распределение – значит случайно распределены данные
  Если не нормально – чаще всего есть причина!!!



Есть дополнительный фактор!!!!! Надо учитывать

Такого быть не может. Подтасовка данных



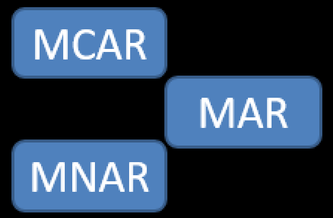
Разведочный анализ данных

2. Пропуски.

Пропуски.

В 1976 году математик Дональд Рубин (Donald B. Rubin)

1. Полностью случайные пропуски (missing completely at random, MCAR)
2. Случайные пропуски (missing at random, MAR)
3. Неслучайные пропуски (missing not at random, MNAR)



Вероятность пропуска зависит от некоторого известного нам фактора.

Пример: Одна часть респондентов (например девушки) лучше отвечают на какой-то вопрос, чем другая часть респондентов (юноши2)

Вероятность пропусков всегда одинакова, так как эти пропуски не связаны с данными.

Пример1: упал прибор и разбился. Измерения не занесли в таблицу

Вероятность появления пропуска зависит от фактора. Просто мы о нём не знаем.

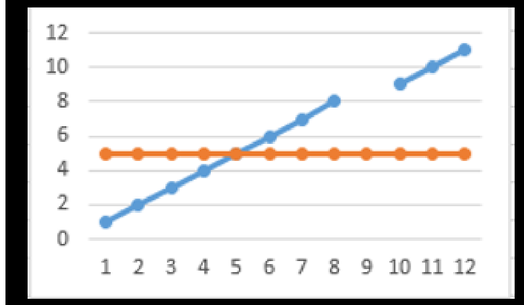
Пример: шкала прибора ограничивается пределом. А система настроена так, что не записывает значение вне предела измерительной системы

Разведочный анализ данных

2. Пропуски.

Пропуски. Количественные данные

Где пропуск



Чем заполнять

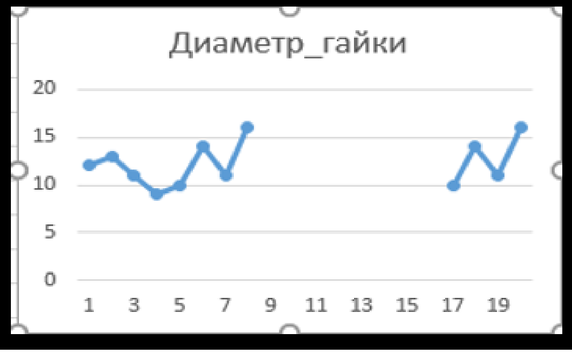
1. Среднее/медиана
2. Предыдущее (среденее из n-значений)/ последующее.
3. Случайное число из интервала
4. Интерполяция
5. Удалить

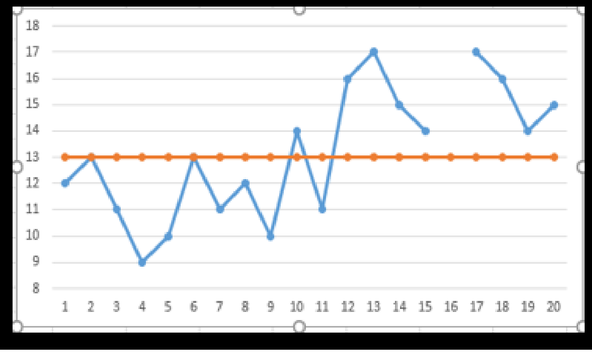
КОМАНДА В ПИТОНЕ

1. median()
2. fillna(metod = ‘ffill’/’bfill’)
   Можно посмотреть SimpleImputer
3. Interpolate(method = ‘linear’).

В документации есть и другие методы.

4. dropna()



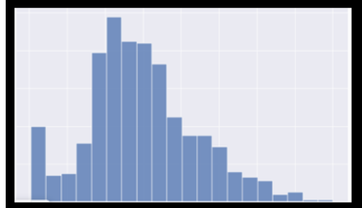


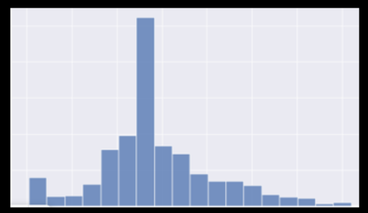
Разведочный анализ данных

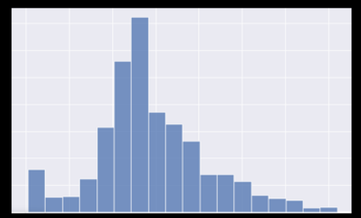
2. Пропуски.

Пропуски. Количественные данные

ШАГИ	КОМАНДА В ПИТОНЕ	ГРАФИК
1. Есть ли пропуски 2. Мало/много/периоды пропусков 3. Какой характер пропусков 4. Принятие решения	df.isna().sum(), df.isna().mean() Встроенной функции расчёта не встречал	import missingno as msno msno.bar(df) График линии, динамика во времени







Пропуски. Категориальные данные

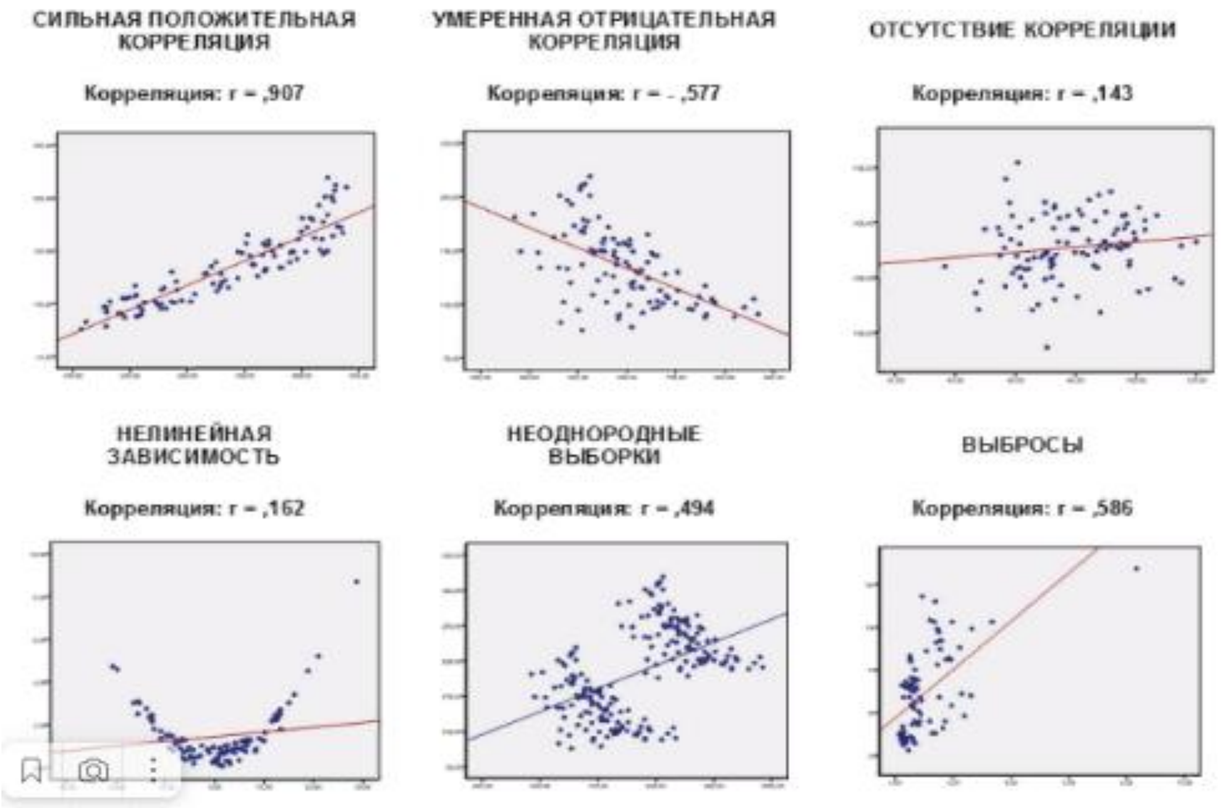
1. Мало пропусков
   - Заполняем модой (наиболее часто встречаемое значение)
1. Много пропусков
   - Ввести новую категорию: «неизвестное»
   - Удалить

Разведочный анализ данных

3. Взаимосвязи

Чаще всего это:

- не линейность связи
- выбросы
- бимодальность



Разведочный анализ данных

4. Дублирование

Одна и та же запись встречается несколько раз

5. Указание размерности

Да и в целом – фиксирование понятийного аппарата у всей аналитической команды – важнейший момент. Он позволяет избежать многих не точностей при решении задач. И экономит время на исправление этих неточностей. Самый известный пример, который гуляет в инете на эту тему: это история про спутник, который конструировали американские и английские инженера. Но одни использовали для измерения импульса «ньютон-секунда», а другие «фунт-секунда»

6. Ещё один важный момент

И такой организационный момент. Он не всегда зависит от специалиста, а больше от менеджмента. Это касается фиксирования истории «касания» исполнителей с данными. Кто и как выгружал? кто и как преобразовывал? и т.д. Для чего? Это помогает быстрее устранять системные ошибки работы над с качеством данных в целом.

Практическая реализация предобработки данных.

Что такое предобработка данных?

Не качественные данные → Необходимо преобразование

Внешний вид загруженных данных

- Название столбцов (пробелы, символы, заглавные буквы..)
- Пропуски
- Название категорий (пробелы, символы, заглавные буквы..)
- Дублирование

Суть исходных данных

- Выбросы
- Адекватность данных бизнесу

Практическая реализация предобработки данных.

Что такое предобработка ?

Работа с названиями столбцов

- df.columns – вывод столбцов
- df.columns.str.lower() – убрать заглавные буквы из названий
- df.columns.str.lower().str.replace(‘ ‘, ‘_’) – убрать заглавные буквы из названий и пробелы в названиях

Работа с дубликатами строк

- df.duplicated() – вывод булевого значений повторяющихся строк
- df.duplicated().sum() - сумма дубль-строк
- df[df.duplicated() == True] – вывод самих строк, чтобы проанализировать (всех)
- df[df.duplicated(keep == ‘first’] – вывод самих строк, чтобы проанализировать (только дублей)
- df.drop_duplicated() - удаление дублей

Что такое предобработка ?

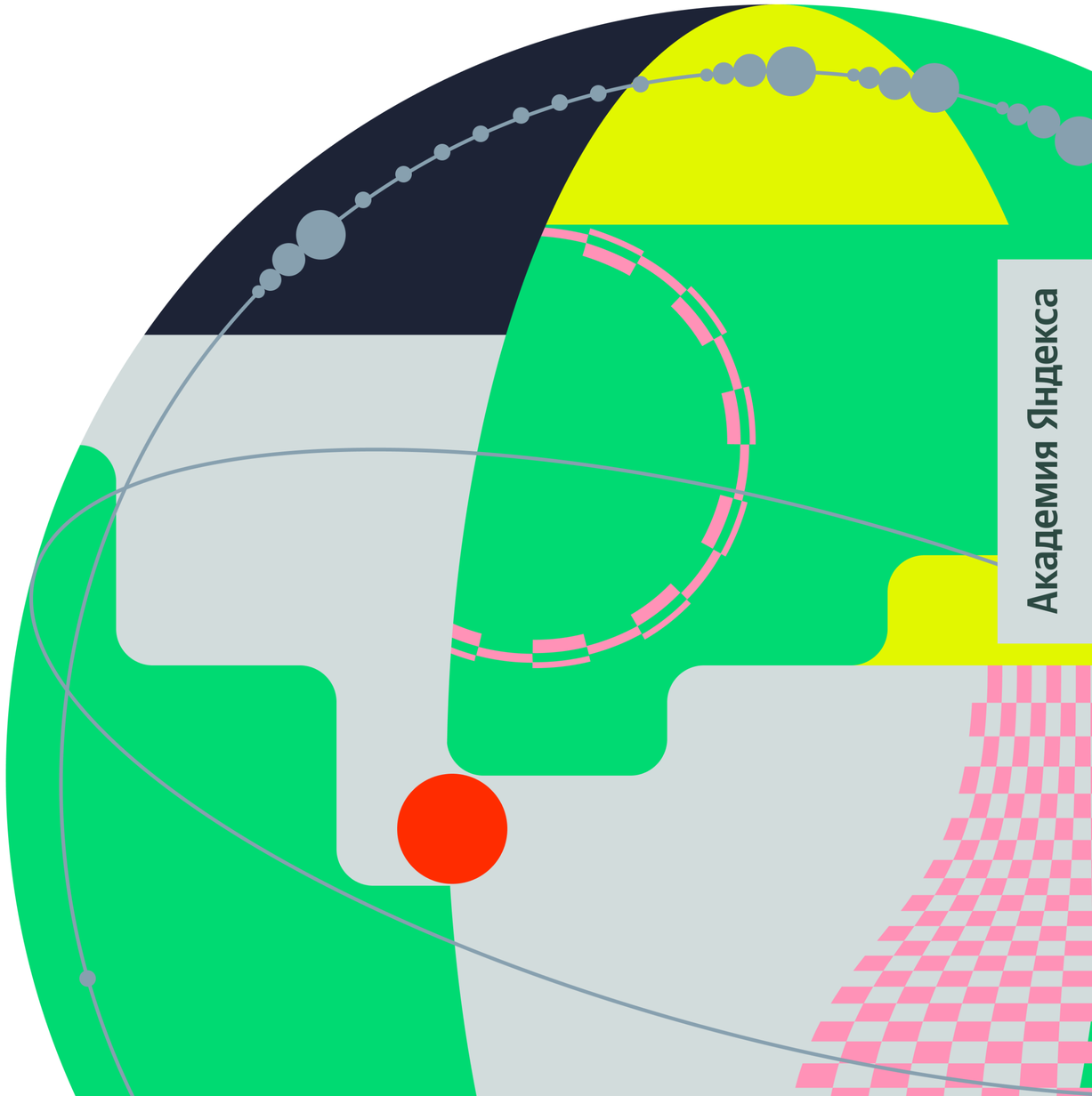
Работа с названиями категорий

- df[‘курение].value_counts(dropna = False) – вывести количество строк по категориям с учётом пропусков
- df[‘курение].unique() - вывод уникальных значений категории.
- df[‘курение] = df[‘курение].map(Словарь) – способ привести названия категорий к нужному виду

Работа с пропусками

- df.isna().mean() – вывод количества пропусков
- df[‘курение].fillna(-1, inplace = True) – способ привести названия категорий к нужному виду
- df.dropna(inplace = True) – удалили все строки с пропусками

Спасибо за внимание



Академия Яндекса позволяет школьникам и студентам освоить востребованные IT-профессии по программам, разработанным экспертами компании