

Первичный и визуальный анализ данных

Урок 1

На этой лекции вы найдете ответы на такие вопросы как:

С чего мы начинаем работу над проектом?

Как проводить первичный анализ?

Для чего нужен визуальный анализ?



Булгакова Татьяна




Преподаватель в GeekBrains, Нетология, Skillfactory

С 2010 года занимаюсь DataScience и NN. Фрилансер

- Участвовала в разработке программы по настройке оборудования для исследования пространственного слуха китообразных НИИ ИПЭЭ РАН
- Участвую в разработке рекомендательных систем по настройке нейростимуляторов для медицинских центров
- Работаю над курсом по нейронным сетям



Что будет на уроке сегодня

-  Мы узнаем, с чего мы начинаем работу над проектом?
-  Как проводить первичный анализ?
-  Для чего нужен визуальный анализ?

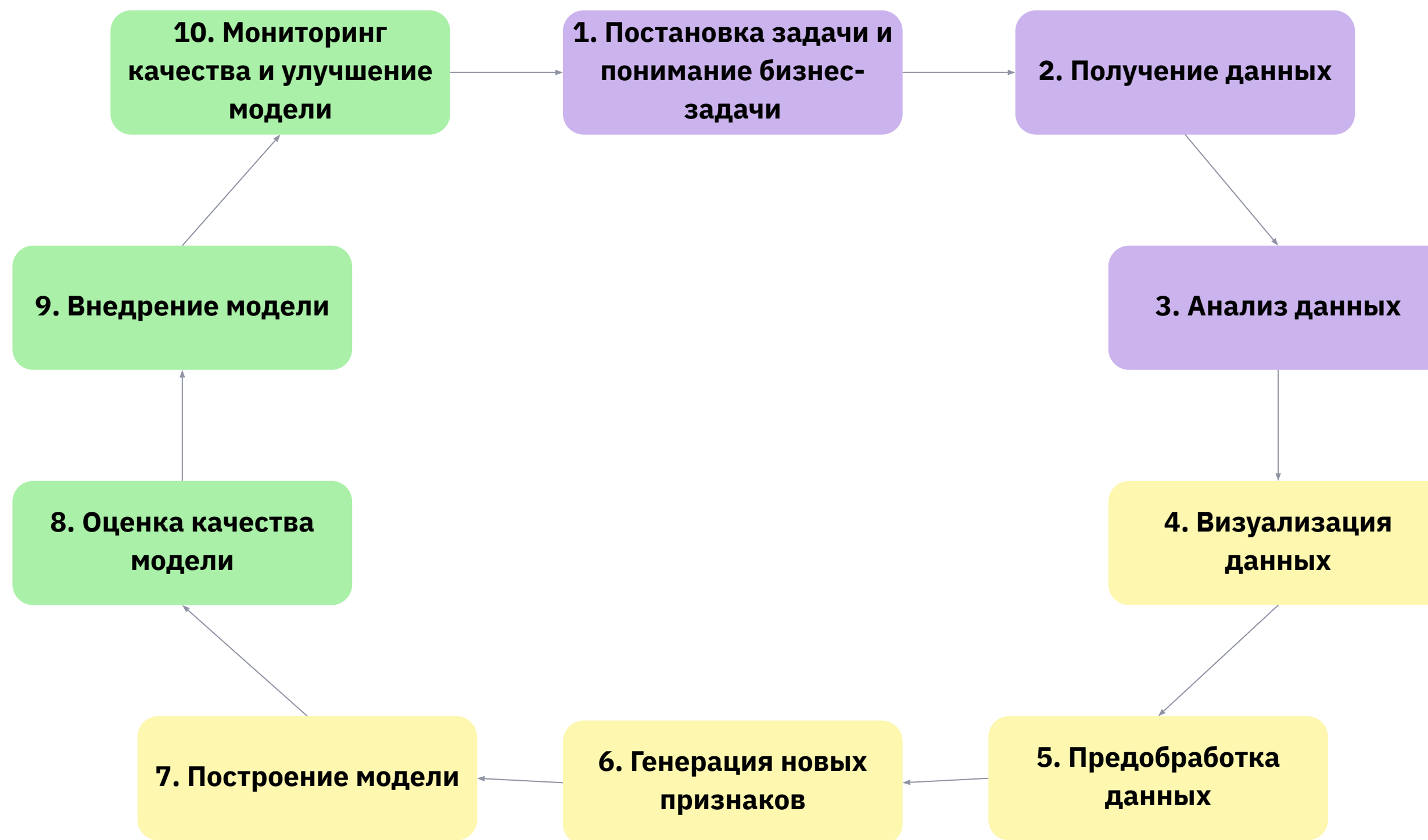


Data Science проект

это прикладные исследования, состоит из таких этапов,
как формулировка гипотез, проектирование
экспериментов и, конечно, оценка результатов и их
пригодности для решения конкретных случаев.



Этапы работы над Data Science проектом:





На что стоит обратить внимание!



Соберите все имеющиеся в вашем распоряжении данные. Вы не знаете точно, какие данные вам понадобятся, и у вас может быть только один шанс собрать их.



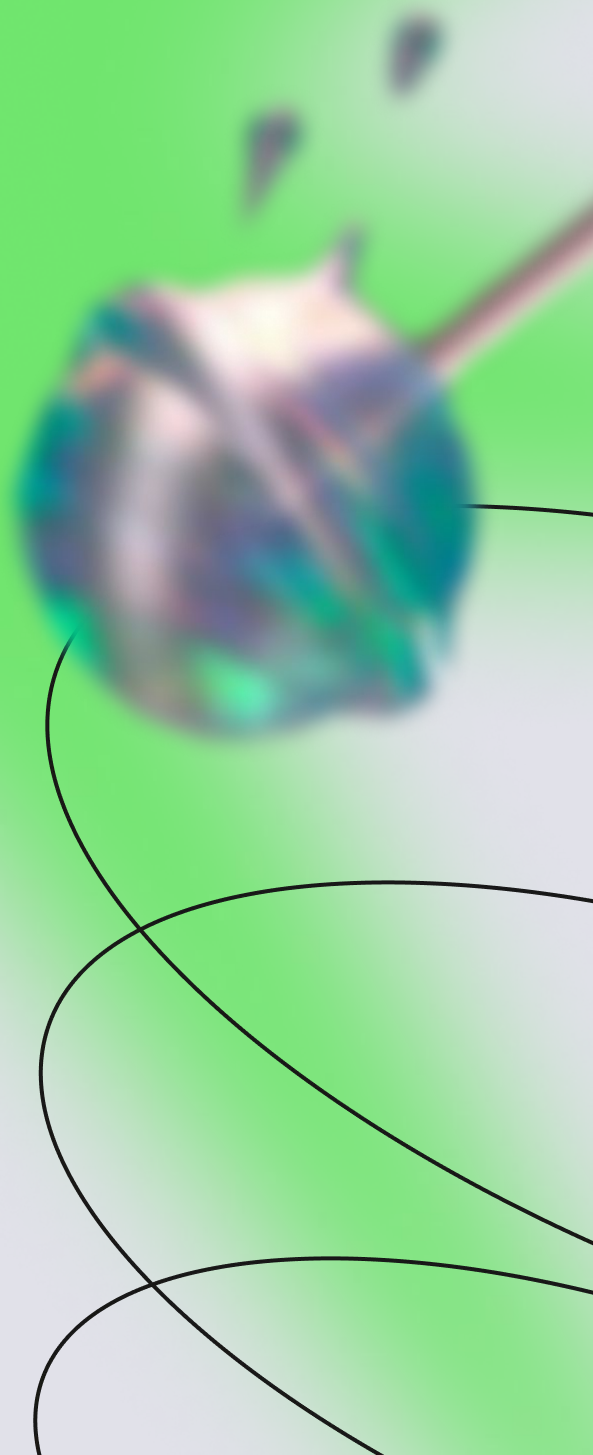
Несколько источников данных дают более полную картину и позволяют лучше оценить поведение пользователей в контексте, но обратной стороной является разнообразие форматов и стоимость их интеграции в аналитическую систему.



Если у вас много данных, можете ли вы обеспечить их хранение и обработку? Ценные данные не всегда доступны в больших объемах, и наоборот.



Первичный
анализ





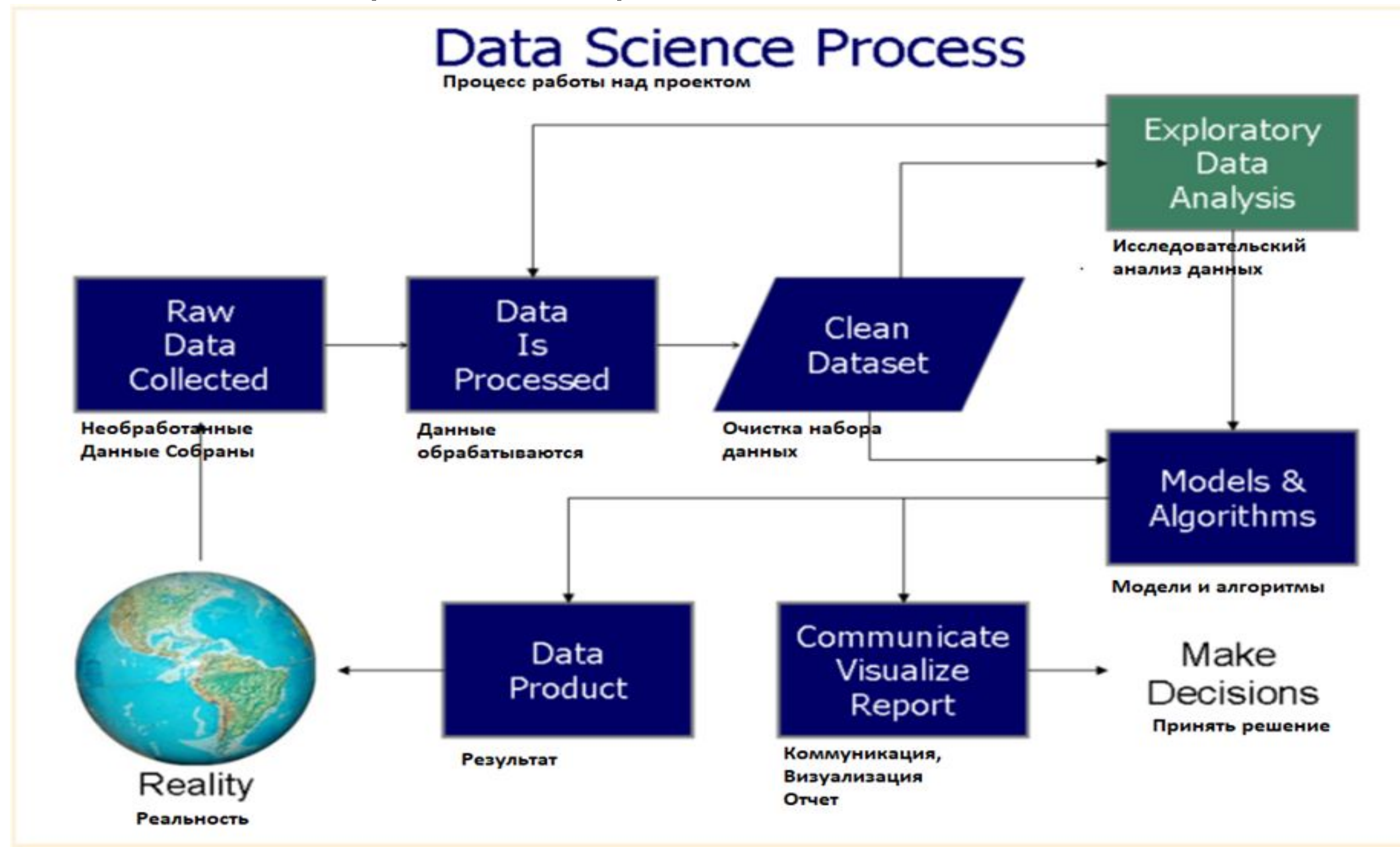
Исследовательский анализ данных

это процесс анализа или понимания данных и извлечения идей или основных характеристик данных. EDA обычно подразделяют на два метода, т.е. визуальный анализ и первичный анализ.











Прежде чем перейти к рассмотрению EDA, важно понять, как EDA вписывается в общий процесс обработки данных.





Технически основной целью EDA является:

-  Изучение распределение данных
-  Обработка отсутствующих значений набора данных (наиболее распространенная проблема с каждым набором данных)
-  Обработка выбросов
-  Удаление повторяющихся данных
-  Кодирование категориальных переменных
-  Нормализация и масштабирование



EDA предполагает решение трех основных задач



Описание данных



Поиск различий



Выявление закономерностей.



Описание данных

- Описание данных предполагает одномерный анализ, поскольку одновременно рассматривается только один атрибут. Для категориальных данных сначала нужно найти уникальные категории, а затем оценить количество в каждой категории.
- При анализе количественных переменных мы оцениваем меры дисперсии для средних, стандартных отклонений, диапазонов, процентилей и других реальных переменных.



Поиск различий

- Поиск различий — это многомерный анализ, поскольку в нем участвует более одной переменной. В частности, можно обнаружить различия по одному признаку под влиянием другого признака.
- Для количественных атрибутов также можно найти различия, сосредоточившись на конкретных категориях.

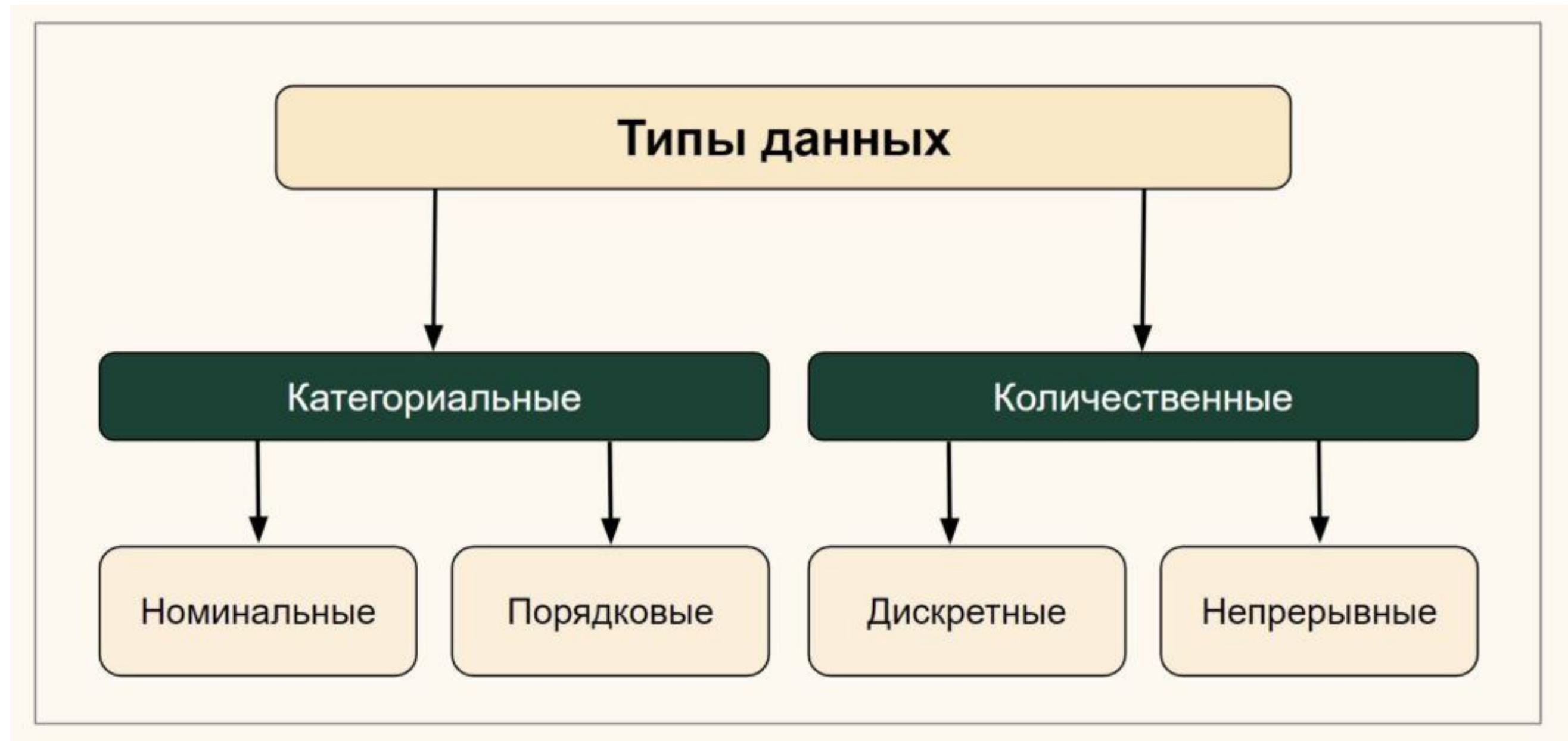


Выявление закономерностей

- Закономерности или взаимосвязи в данных могут быть выявлены между двумя количественными признаками



Вспомним какие данные бывают:





Описание данных:

Задача: **описание данных**

Категориальные данные

- `.unique()` и `.value_counts()`
- `df.describe()`
- `barplot`
- `countplot`

Количественные данные

- `df.describe()`
- гистограмма
- график плотности
- `boxplot`
- гистограмма + `boxplot`



Типы пропусков:



Полностью случайные пропуски предполагают, что вероятность появления пропуска никак не связана с данными. Такие пропуски возникают, например, если измерительный прибор неисправен и случайным образом не записал часть наблюдений, или если один из образцов крови, изучаемых в лаборатории, оказался поврежден и по этой причине его характеристики выпали из исследования.



Случайные пропуски — вероятность появления пропуска зависит от некоторой известной нам переменной. Например, отсутствие ответа на определенный вопрос анкеты может зависеть от возраста респондента.





Неслучайные пропуски — вероятность появления пропуска зависит, в том числе, от фактора, о котором мы ничего не знаем. Например, у весов может быть верхний предел измерения и любой образец выше этого предела автоматически не записывается.



При наличии пропусков нужно решить, что можно с ними сделать:

 Можно удалить, в том случае, если пропусков немного относительно всей выборки.

 Если пропусков много (более 50% от общего количества объектов), то, вероятно, от признака стоит избавиться.

 Пропуски можно заменить медианой или модой (в случае категориальных признаков). Медиана более стабильна нежели среднее арифметическое. Выбросы оказывают сильное влияние на среднее арифметическое.

 В ряде случаев пропуски можно заменить, используя взаимное влияние, то есть, с учетом других признаков восстановить пропуски в нужном признаке.

 Можно использовать модель для восстановления пропущенных значений.



Одномерный и многомерный анализ:

В соответствии с еще одной классификацией данные предполагают одномерный и многомерный анализ.



При **одномерном анализе** (univariate analysis) мы сосредоточены на изучении одного единственного показателя. **Многомерный анализ** (multivariate analysis) предполагает, что мы изучаем сразу несколько признаков.



Выводы:

1

Одномерный анализ и Многомерный анализ

Одномерный анализ - это анализ переменных по отдельности.

Если переменная анализируется отдельно от других переменных, категориальных или непрерывных, это называется одномерным анализом.

Многомерный анализ предполагает, что мы изучаем сразу несколько признаков.

2

Визуализация

Визуализация - это мощный инструмент для донесения мыслей и идей до конечных пользователей, а также помощь в восприятии и анализе данных.

3

EDA

EDA - один из самых важных этапов проекта по науке о данных. Этот этап не только определяет направление проекта, но и помогает максимально эффективно использовать набор данных.



Спасибо за внимание

