1. **Способы обнаружения пропущенных данных:**
   * Визуальный осмотр данных: это может быть неэффективно для больших наборов данных.
   * Использование функций языка программирования, таких как isnull(), notnull(), info() в Python с библиотекой pandas.
   * Визуализация данных с помощью тепловых карт, где пропущенные данные обозначаются отличным цветом.
2. **Определение типа данных каждого признака:**
   * В Python с библиотекой pandas можно использовать функцию dtypes или info(), чтобы увидеть типы данных каждого столбца.
3. **Пример категориальных данных:**
   * Пол (мужской, женский)
   * Тип автомобиля (седан, хэтчбек, внедорожник)
   * Город (Москва, Санкт-Петербург, Новосибирск)
4. **Способы кодирования категориальных данных:**
   * **One-Hot Encoding:** каждая категория становится отдельным столбцом, где 1 обозначает наличие категории, а 0 - отсутствие.
   * **Label Encoding:** каждой категории присваивается уникальное числовое значение.
   * **Ordinal Encoding:** используется для порядковых категориальных данных, где категории имеют определенный порядок.
5. **Как работает One-Hot Encoding:**
   * One-Hot Encoding преобразует каждую категорию в отдельный столбец. Если у нас есть категориальный признак с N категориями, то после One-Hot Encoding мы получим N столбцов.
   * Каждый из этих столбцов будет бинарным, то есть содержать только 0 или 1. Значение 1 обозначает наличие соответствующей категории, а 0 - отсутствие.
6. **Другие ошибки данных, помимо пропусков и выбросов:**
   * Неконсистентные данные: например, использование разных единиц измерения.
   * Дубликаты: повторяющиеся строки в данных.
   * Неправильные значения: например, отрицательный возраст или вес.
   * Нерелевантные данные: данные, которые не имеют отношения к анализу.
   * Неправильная структура данных: например, столбцы и строки перепутаны местами.
   * Типы данных не соответствуют ожидаемым: например, числовые значения в текстовом поле.
   * Ошибки ввода: например, опечатки или неправильное использование регистра.
   * Несбалансированные данные: в данных для классификации один класс значительно преобладает над другими.