**Вопросы к лабораторной работе №2**

1. Перечислите различные способы обнаружения пропущенных данных.

- Визуальный анализ данных с помощью графиков (тепловая карта корреляции, диаграмма рассеяния, гистограммы).

- Использование методов pandas, таких как `.isnull()` и `.notnull()` для выявления пропущенных значений.

- Использование методов визуализации данных, таких как seaborn.heatmap() для визуализации пропущенных значений.

- Расчет общего количества пропущенных значений для каждого признака.

2. Как можно определить тип данных каждого признака?

- Использование метода dtypes в pandas DataFrame.

- Использование метода info() в pandas DataFrame.

- Использование метода describe() в pandas DataFrame, который также показывает статистическую информацию о числовых признаках.

3. Приведите пример категориальных данных

- Пол (мужской, женский)

- Тип товара (одежда, обувь, аксессуары)

- Уровень образования (высшее, среднее, начальное)

- Раса (белый, черный, азиатский)

4. Какими способами можно закодировать категориальные данные?

- Label Encoding: Замена категориальных значений числовыми метками (0, 1, 2, ...).

- One-Hot Encoding: Создание бинарных (дамми) переменных для каждой категории.

- Ordinal Encoding: Присвоение числовых значений категориям с учетом их порядка.

- Target Encoding: Преобразование категориальных значений в средние значения целевой переменной.

5. Как работает One-Hot Encoding?

Для каждого уникального значения категориальной переменной создается новый бинарный признак. Этот новый признак равен 1, если категориальный признак имеет соответствующее значение, и 0 в противном случае.

Это позволяет сохранить информацию о принадлежности к определенной категории, не устанавливая ложные отношения порядка между категориями.

6. Какие еще встречаются ошибки данных помимо пропусков и выбросов?

- Дубликаты данных: повторяющиеся записи в датасете.

- Неверные типы данных: неправильно определенные типы данных для признаков (например, строковые значения, которые должны быть числами).

- Несоответствие форматов данных: некорректный формат данных для определенных признаков (например, дата и время записаны как строки).

- Искаженные данные: данные, которые содержат ошибки или неточности в силу человеческого фактора или технических проблем.