

Задание на разведочный анализ данных для датасета `diamonds.csv`:

Цели задания:

1. Понять основные статистические характеристики датасета.
2. Научиться визуализировать данные для анализа.
3. Исследовать взаимосвязь между различными переменными.

Описание датасета:

Датасет `diamonds.csv` содержит следующие столбцы:

- `carat`: вес алмаза в каратах.
- `cut`: качество огранки алмаза (Fair, Good, Very Good, Premium, Ideal).
- `color`: цвет алмаза, от J (худший) до D (лучший).
- `clarity`: чистота алмаза (от I1 до IF).
- `depth`: общая глубина в процентах от вершины до основания.
- `table`: ширина верхней части алмаза относительно самой широкой точки.
- `price`: цена алмаза в долларах США.
- `x`: длина алмаза в мм.
- `y`: ширина алмаза в мм.
- `z`: глубина алмаза в мм.

Задачи:

0. Очистка датасета

1. Исследование мер центральной тенденции:

- Найдите среднее значение, медиану и моду для численных переменных.
- Опишите, что каждая из этих мер говорит о распределении данных.

2. Визуализация данных:

- Создайте гистограммы для переменных `carat`, `depth`, и `price`.
- Постройте boxplot для цены в зависимости от качества огранки и цвета алмазов.
- Используйте scatter plot для изучения взаимосвязи между `carat` и `price`.

3. Исследование мер вариативности:

- Рассчитайте стандартное отклонение и диапазон для переменных `carat`, `depth`, и `price`.
- Объясните значение этих мер в контексте данных.

4. Корреляционная матрица:

- Создайте корреляционную матрицу для всех численных переменных.
- Используйте тепловую карту для визуализации корреляционной матрицы.
- Обсудите, какие переменные наиболее сильно коррелируют и предположите почему.