Разведочный анализ данных

Разведочный анализ данных (Exploratory Data Analysis, EDA) — это начальный этап работы с данными, направленный на изучение структуры, характеристик и основных свойств набора данных. Его цель — лучше понять, с какими данными предстоит работать, выявить закономерности и аномалии, а также подготовить данные для дальнейшего анализа или построения моделей.

Основные задачи и этапы EDA:

- понимание структуры данных: изучение объёма данных, типов переменных (числовые, категориальные), наличия пропущенных значений и дубликатов;
- Анализ распределений переменных: оценка распределения каждого признака, поиск выбросов и аномалий;
- выявление взаимосвязей: исследование корреляций и зависимости между признаками;
- обнаружение аномалий и выбросов: нахождение необычных наблюдений, которые могут искажать анализ;
- проверка гипотез: формулировка и первичная проверка предположений о данных;
- подготовка данных: чистка (удаление или заполнение пропусков), преобразование, масштабирование и прочие операции предобработки.

Инструменты и методы EDA часто включают статистические сводки (средние, медианы, стандартные отклонения), таблицы частот, визуализацию (гистограммы, ящики с усами — boxplots, диаграммы разброса), корреляционные матрицы, тепловые карты и другие графические средства.

EDA помогает исследователю «проникнуть» в данные, что позволяет выявить важные характеристики и подготовить адекватные модели или принять обоснованные решения при анализе.

Таким образом, EDA — это фундаментальный исследовательский этап в работе с данными, необходимый для понимания, качественной предобработки и построения надежных аналитических моделей.

Обнаружение и устранение аномалий (выбросов, некорректных или необычных значений) — важный этап разведочного анализа данных (EDA). Основными шагами являются обнаружение аномалий при помощи инструментов визуализации, статистические методы.

Часто используются такие статистические методы:

- правила на основе z-оценки (Z-score) — значения с абсолютным z-score выше порогового значения (обычно 3) могут считаться аномалиями;

- интерквартильный размах (IQR) — выбросы, выходящие за границы IQR, считаются потенциально аномальными;

- метрики плотности и расстояний — например, метод локальной плотности (Local Outlier Factor, LOF), расстояния до k-ближайших соседей.

Устранение и обработка аномалий. Удаление — если выбросы являются ошибками или нерепрезентативными точками, их можно удалить. Однако нужно понимать, что иногда выбросы — важные данные. Замена выбросов может производиться при помощи медианы, среднего или наиболее частым значением. Более подробно замена аномальных значений и заполнение пропусков описаны в отдельном документе. Также может использоваться преобразования, например, логарифмирование, нормализация, стандартизация для смягчения влияния экстремальных значений на модель.

Отдельная обработка — выделение аномалий в отдельный класс (например, в задачах классификации) и использование специализированных алгоритмов.

Для разведочного анализа данных (EDA) основные методы визуализации помогают быстро понять структуру, распределение и взаимосвязи в данных. Они служат для выявления закономерностей, аномалий, трендов и проблем в данных на этапе предварительного анализа.

Для визуализации аномалий часто используются:

- Ящик с усами (boxplot) — показывает медиану, квартильные границы и выбросы по формуле интерквартильного размаха (IQR) (рис.1). Выбросы — точки за пределами интервала [Q1 - 1.5 * IQR, Q3 + 1.5*IQR].

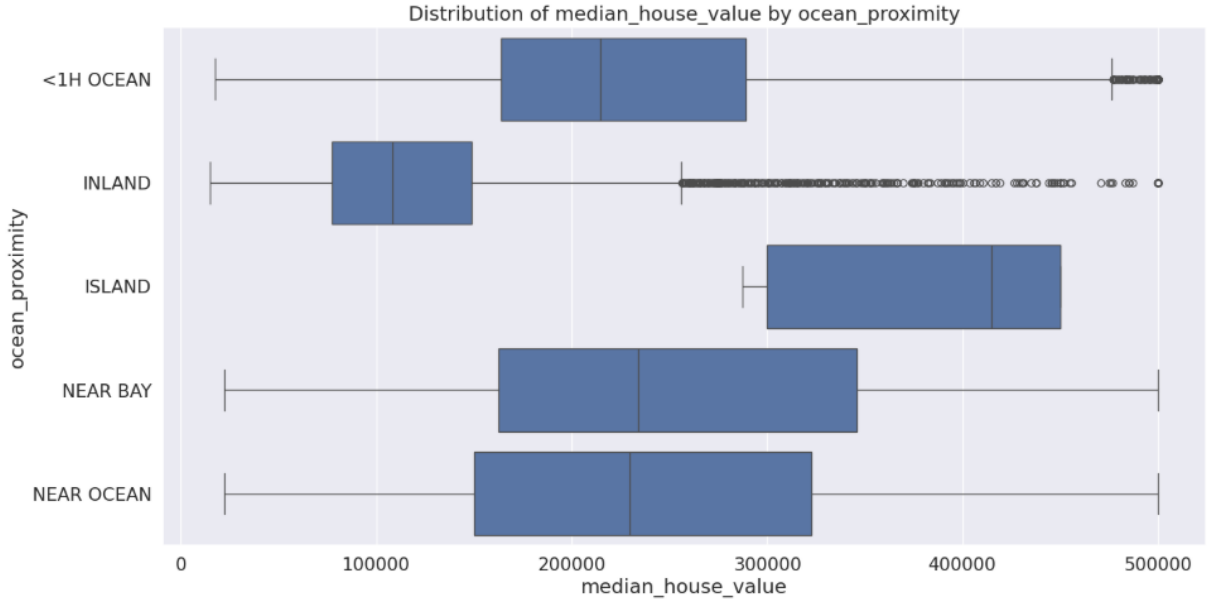

Рисунок 1 – Выявление аномалий про помощи boxplot

- Гистограмма — даёт представление о распределении данных и выявляет экстремальные значения или необычные пики (рис.2).

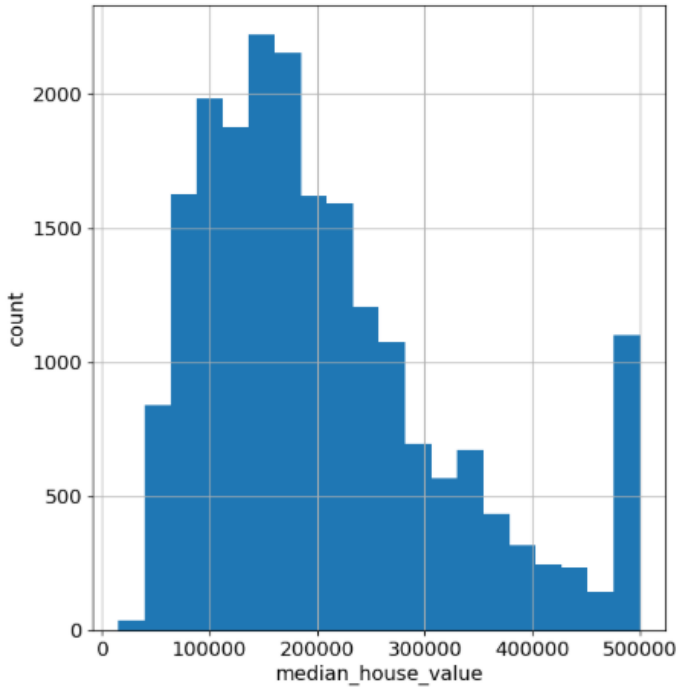

Рисунок 2 - Выявление аномалий при помощи гистограммы

Диаграммы рассеяния (scatter plot) — помогают визуально выявить необычные точки в двухмерных признаках (рис.3).

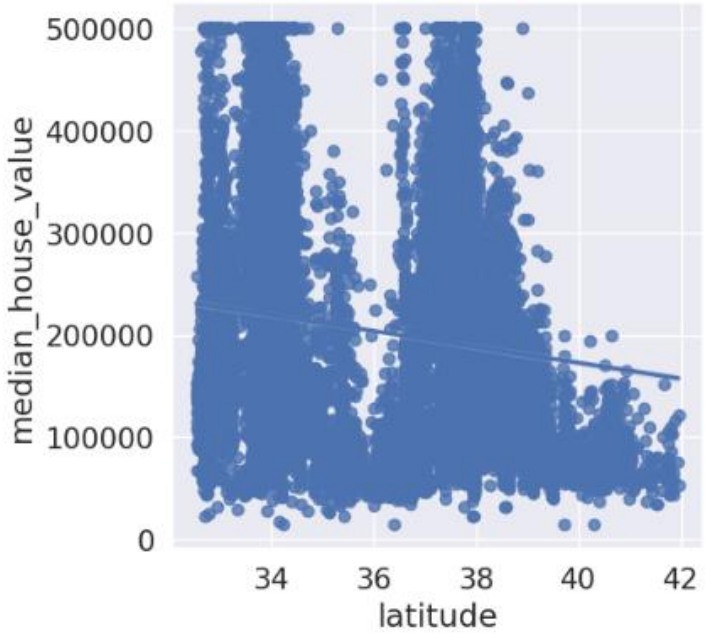

Рисунок 3 – Выявление закономерностей и выбросов при помощи диаграммы рассеяния

Также применяют такие диаграммы как:

- столбчатая диаграмма (Bar chart). Подходит для сравнения категориальных данных или частот. Позволяет визуально оценить количество элементов в каждой категории;

- круговая диаграмма (Pie chart). Используется для представления долей частей в целом. Практична для небольшого количества категорий, чтобы быстро показать пропорции;

- тепловая карта (Heatmap). Помогает визуализировать матрицу данных, например, корреляции между переменными с помощью цветового градиента. Полезна для обнаружения сильных связей и паттернов;

- линейный график (Line plot). Применяется для анализа динамики и трендов во времени или по порядку наблюдений;

- парные диаграммы (Pairplot). Показывают матрицу диаграмм рассеяния и распределений для нескольких переменных сразу, что удобно для поиска взаимосвязей в многомерных данных.

Визуализация данных в EDA позволяет упрощать и ускорять интерпретацию данных, обнаруживать закономерности, тренды, аномалии и выбросы, проверять качество и полноту данных, делать выводы и принимать решения о дальнейшем анализе и предобработке, эффективно коммуницировать результаты анализа заинтересованным лицам.