Разведочный анализ данных (EDA)

Что нужно узнать о таблице, прежде чем делать выводы и обучать модель?

Вы записали несколько поездок и хотите оценить время дороги. Одна длилась сто минут, у другой время потерялось. Можно ли сразу взять среднее? Исследуем небольшую таблицу и проследим, как решения при обработке меняют вывод.

1. Одна строка — одна поездка

Представьте, что вы неделю записывали дорогу до школы и теперь хотите понять, сколько времени закладывать. Сначала нужно разобраться, что именно записано. Одна строка нашей таблицы описывает одну поездку; столбцы хранят её номер, расстояние и длительность. Это учебный пример с придуманными наблюдениями, поэтому выводы относятся только к нему.

Поездка	Расстояние, км	Время, мин
А	2	10
Б	3	12
В	3	12
Г	4	16
Д	5	100
Е	2	—

Номер поездки помогает различать записи. Он не измеряет величину: нельзя осмысленно усреднить буквы А и Б. Расстояние и время, напротив, числовые признаки. Единицы важны: число 100 может означать минуты, секунды или ошибку ввода. До вычислений уточните, когда и каким способом измерялось время. Включены ли ожидание автобуса и остановки по дороге?

Разведочный анализ данных, или EDA, начинается с таких вопросов. Мы выясняем устройство набора, проверяем качество записей и ищем заметные закономерности. Результат исследования должен объяснять, откуда взялось каждое число. Вычислить среднее, не поняв смысл строк, можно технически правильно и содержательно бессмысленно.

Пример

Если приложение сохраняло положение каждую минуту, строка описывала бы момент наблюдения, а не отдельную поездку. Среднее по таким строкам сильнее учитывало бы длинные поездки.

> В журнале 8 поездок, из них две без времени. Сколько поездок имеют известную длительность?

2. Пропуски и похожие записи

У поездки Е длительность неизвестна. Прочерк означает отсутствие измерения; ноль означал бы измеренную поездку нулевой продолжительности. Подставить ноль вместо прочерка — значит

придумать наблюдение. При положительных известных длительностях такая подстановка уменьшит среднее.

В исходных шести строках один пропуск времени. Доля пропусков равна одной шестой, около 16,7 процента. Но среднее по известным временам использует только пять значений. В отчёте должны быть названы оба числа: всего шесть поездок, время известно для пяти. Иначе читатель может решить, что результат учитывает весь журнал.

Поездки Б и В имеют одинаковые расстояние и время. Этого недостаточно для удаления одной строки. Человек мог дважды проехать один маршрут с одинаковой скоростью. **Дубликат** — повторная запись того же наблюдения, а не просто совпадение нескольких признаков. Чтобы обнаружить случайную повторную загрузку, понадобятся дата, идентификатор и правило записи.

Удаление строк тоже меняет исследуемую группу. Если время чаще не записывается при разрядившемся телефоне во время долгой дороги, оставшиеся поездки будут смещены в сторону коротких. Сначала опишите пропуски и возможную причину, затем выбирайте обработку. Неизвестная причина должна остаться ограничением вашего вывода.

Заметка
Сохраняйте исходную таблицу. Исправления ведите отдельной копией с пояснением: что изменили, почему и сколько строк это затронуло. Тогда другой человек сможет повторить расчёт.

> В журнале 12 поездок. У трёх время неизвестно; сумма известных времён равна 180 минутам. Чему равно среднее по известным временам?

Профи. **Пропуск в коде: `None`, а не ноль**

В таблице пропуск обозначен прочерком; в Python для него есть значение `None` — «нет значения». Список времён тогда выглядит так: `[10, 12, 12, 16, 100, None]`. Команда `sum` на таком списке остановится с `TypeError`: складывать число с `None` нельзя. Это полезно: программа отказывается считать, пока вы не решите, что делать с пропуском.

Решения бывают разные, и у каждого свой след. Замена `None` нулём сохраняет шесть строк и даёт среднее $150/6 = 25$ вместо 30: придуманное наблюдение утянуло среднее вниз. Замена средним известных, 30, сохраняет среднее, зато уменьшает разброс: шесть значений выглядят более похожими друг на друга, чем пять измеренных, и любая оценка надёжности вывода станет завышенной. Исключение строки оставляет 30, но меняет размер группы, и это нужно сказать вслух: «среднее по пяти из шести».

Как поймать ошибку: печатайте число учтённых значений рядом со средним. Если `len(known)` равно числу строк таблицы, хотя пропуски были, значит их чем-то заполнили. В упражнении стартовый код заменяет пропуск нулём; перепишите строку так, чтобы пропуск был исключён, и убедитесь, что известных значений пять.

Исключить пропуск

Цель: Оставьте в known только измеренные значения, без None.

```
times = [10, 12, 12, 16, 100, None]
known = [t if t is not None else 0 for t in times]  # TODO: не заменять,
а исключить
average = sum(known) / len(known)
print(f"Учтено {len(known)} значений, среднее {average:.1f} мин")
```

3. Среднее и середина ряда

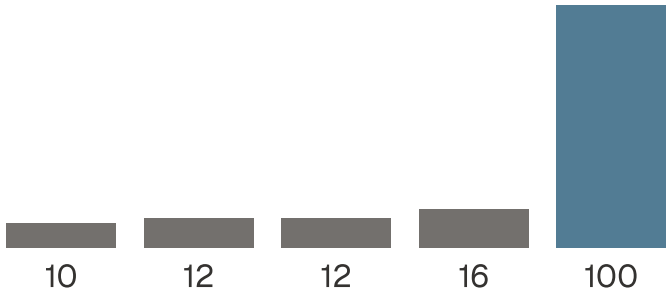
Для пяти известных времён сумма равна 150 минутам. **Среднее** получается делением суммы на количество наблюдений: 150 делим на 5 и получаем 30 минут. Это величина, которую имела бы каждая поездка при равномерном распределении общего времени между ними.

Однако четыре поездки заняли не больше шестнадцати минут. Число 30 плохо описывает их обычную длительность. Посмотрим на **медиану**: упорядочим значения и возьмём середину. Ряд 10, 12, 12, 16, 100 уже упорядочен. Посередине стоит 12. Если значений чётное число, берём среднее двух центральных: для 10, 12, 16, 20 медиана равна 14.

Среднее чувствительно к величине каждого наблюдения. Медиана зависит от положения середины ряда. Ни одна характеристика не заменяет весь набор: пять одинаковых поездок по двенадцать минут и наш журнал имеют одну медиану, но совершенно разную предсказуемость времени.

В опыте измените пятую длительность со 100 до 20 минут. Сначала оцените направление изменения среднего. Затем сравните показания. При 100 минутах среднее равно 30, при 20 минутах равно 14; медиана в обоих случаях 12. Здесь каждый столбик показывает одну поездку. Это диаграмма отдельных наблюдений, не гистограмма частот.

Проверьте свой прогноз



Пять поездок: 10, 12, 12, 16, 100 мин. Среднее: 30,0 мин; медиана: 12 мин. Каждый столбик — одна поездка.

Измените пятую поездку и сравните среднее с медианой. Остальные четыре наблюдения сохраняются.

Интерактивный эксперимент — в электронной версии урока.

> Время четырёх поездок равно 8, 10, 18 и 40 минутам. Найдите медиану длительности.

Профи. **Сдвиг среднего: на сколько и почему**

Среднее пяти времён равно 30, потому что сумма равна 150. Если одно значение изменить на Δ, сумма изменится ровно на Δ, а среднее — на Δ/n. Замена 100 на 20 даёт $\Delta = -80$, $n = 5$, сдвиг -16: с 30 до 14. В этом и причина чувствительности среднего: каждое наблюдение входит в него с весом $1/n$, и одно большое отклонение переносится в результат целиком, лишь поделённое на число наблюдений.

Медиана устроена иначе: её задаёт порядок, а не величины. В ряду $10, 12, 12, 16, 100$ последнее значение можно заменить любым числом, и медиана останется 12: два значения 12 держат середину с обеих сторон. Если бы ряд был $10, 11, 12, 16, 100$, медиана 12 держалась бы, пока новое значение не меньше 12; при значении 11,5 она стала бы 11,5, при значении 5 — равной 11. Сдвиг медианы ограничен соседями середины, сдвиг среднего — нет.

Типичная ошибка: исключить 100 как «выброс» и всё равно делить на пять. Сумма оставшихся 50, наблюдений четыре, среднее 12,5, а не 10. Числитель и знаменатель всегда считаются по одной и той же группе; при каждом исключении пересчитывайте оба.

> Среднее пяти времён поездки равно 30 минутам. Одно из них, 100 минут, оказалось опечаткой: верное значение 20 минут. Чему равно исправленное среднее пяти времён? Ответ с точностью до десятых.

Подумай

Если самая длинная поездка действительно заняла сто минут, удалять её только ради красивого среднего нельзя. Она может быть важнее остальных при планировании времени с запасом.

Эпик. **Когда ломается и медиана**

Среднее портит одно наблюдение. Медиану — нет, но и у неё есть предел: она выдерживает ошибки меньше чем в половине записей. Пусть в журнале семь поездок: $10, 11, 12, 13, 14, 15, 16$ минут, медиана 13. Если приложение записало секунды вместо минут в трёх самых длинных поездках, они превратятся в $840, 900, 960$, упорядоченный ряд станет $10, 11, 12, 13, 840, 900, 960$, а медиана останется 13.

Испортите четвёртую запись, и середина ряда окажется среди испорченных. Худший случай для медианы — половина записей, для среднего — одна. Ни одна сводка не защищает от систематической ошибки, затронувшей большинство строк; её замечают по смыслу (дорога в школу длиной 900 минут) и по единицам измерения.

Второй контрпример — двугорбый ряд. Поездки на автобусе занимают $8, 9, 9, 10$ минут, пешком — $28, 29, 30, 31$. Для всех восьми медиана равна 19, среднее — $19{,}25$. Оба числа не описывают ни одной реальной поездки: никто не добирается за девятнадцать минут. Здесь нужна не лучшая сводка, а разделение на две группы; подсказать его может гистограмма или знание о том, что строки неоднородны.

> В журнале семь поездок: 10, 11, 12, 13, 14, 15, 16 минут. Самые длинные поездки по очереди заменяют на ошибочные значения больше 800. Какое наименьшее число замен сделает медиану больше 100?

4. Распределения и связи

Когда наблюдений много, отдельные столбики становятся неудобны. **Гистограмма** объединяет числовую ось в интервалы и показывает, сколько наблюдений попало в каждый. Для наших известных времён в интервале от нуля включительно до двадцати исключительно лежат четыре поездки; от двадцати до ста исключительно нет ни одной; от ста до ста двадцати лежит одна.

Границы интервалов нужно объявлять заранее. Иначе значение ровно на границе может попасть в два столбика или исчезнуть. Узкие интервалы показывают подробности, широкие могут скрыть отдельные группы. Поэтому при чтении гистограммы смотрите на обе оси: диапазон значений и число наблюдений. Большая высота столбца означает частоту, не большую длительность отдельной поездки.

Для двух числовых признаков подходит **диаграмма рассеяния**. Каждая точка показывает одну поездку: по горизонтали расстояние, по вертикали время. Точки А и Г будут иметь координаты (2; 10) и (4; 16). Поездка Д окажется далеко по вертикали. Е нельзя нанести полностью: её вертикальная координата неизвестна.

Связь расстояния со временем ещё не объясняет причину каждой задержки. Время зависит также от транспорта, ожидания и дорожной обстановки. Сравнение разных групп помогает заметить различия, но наличие связи само по себе не доказывает причинный эффект. Для вывода «автобус быстрее велосипеда» понадобились бы сопоставимые маршруты и условия.

Ссылки
Раздел EDA в справочнике NIST объясняет роль графиков в исследовании данных. Численные сводки и графики дополняют друг друга.

Эпик. **Велосипед быстрее на каждом маршруте — и медленнее в целом**

Сравним два способа добираться по журналу за месяц. Короткие маршруты: на автобусе 10 поездок со средним 12 минут, на велосипеде 2 поездки со средним 10 минут. Длинные маршруты: на автобусе 2 поездки со средним 40 минут, на велосипеде 10 поездок со средним 36 минут. На коротких велосипед быстрее ($10 < 12$), на длинных тоже ($36 < 40$). Теперь сложим всё: автобус — $(10 \cdot 12 + 2 \cdot 40)/12 = 200/12 \approx 16{,}7$ минуты, велосипед — $(2 \cdot 10 + 10 \cdot 36)/12 = 380/12 \approx 31{,}7$. В целом велосипед почти вдвое медленнее.

Противоречия нет: среднее по всем поездкам взвешено долями групп, а велосипед чаще выбирали для длинных маршрутов. Правило «посчитай среднее и сравни» ломается, когда у групп разный состав. Это **парадокс Симпсона**: знак различия может поменяться после объединения групп. Чтобы вывод «велосипед быстрее» был честным, сравнивать нужно при одинаковом расстоянии, то есть внутри маршрута, или заранее уравнять составы.

Перенос: та же ловушка возникает при сравнении школ по среднему баллу, больниц по доле выживших и моделей по точности на разных наборах данных. Прежде чем сравнивать два средних, спросите, из каких групп они собраны и совпадают ли доли этих групп.

> Велосипедные поездки за месяц: 2 короткие со средним 10 минут и 10 длинных со средним 36 минут. Чему равно среднее время всех 12 велосипедных поездок? Ответ с точностью до сотых.

5. Вывод, который можно проверить

Хороший итог содержит наблюдение, число и ограничение. Например: «Для пяти поездок с известным временем медиана равна двенадцати минутам. Одно наблюдение равно ста минутам и сильно увеличивает среднее. У шестой поездки время неизвестно; весь журнал слишком мал, чтобы оценить надёжность маршрута на год вперёд».

Сравните это с фразой «обычно дорога быстрая». В ней неясно, что значит «быстрая», какие записи использованы и сколько исключено. Числа не делают вывод автоматически правильным, но позволяют его проверить. Полезно отдельно выписать вопрос: «Поездка Д измерена в тех же единицах и включает те же этапы дороги?» Ответ на него может изменить расчёт.

Ниже можно проверить среднее программой. Список содержит только известные времена. В стартовом коде сумма делится на шесть, как будто пропуск был известным наблюдением. Исправьте знаменатель через длину списка. Код здесь повторяет уже выполненный ручной расчёт и помогает заметить разницу между числом строк и числом измерений.

Проверить знаменатель

Цель: Разделите сумму на число известных времён.

```
times = [10, 12, 12, 16, 100]
average = sum(times) / 6  # TODO: используйте число измерений
print(f"Среднее: {average:.1f} мин")
```

Где это встретится дальше

Сначала выясните, что и как измерено. Затем читайте распределения и связи, сохраняя ограничения наблюдений.

Различить пропуск, ноль и повторную запись.

Посчитать среднее и медиану и объяснить влияние необычного значения.

Сформулировать вывод с числом и ограничением.

История

Джон Тьюки развивал исследовательский подход к анализу данных с опорой на графики. Введение NIST описывает эту традицию. Здесь важен сам приём: рассмотреть наблюдения до выбора модели.

Заметка

Начните отчёт с числа строк и определения одной строки. Затем укажите пропуски, выбранные сводки и необычные наблюдения. Отделяйте проверенный факт от гипотезы о его причине.

Домашняя работа

1. Для ряда 5, 7, 8, 20 найдите среднее и медиану.
2. Придумайте две разные причины неизвестного времени поездки и объясните возможное смещение.
3. ★ Подберите два ряда с одинаковым средним, но заметно разным разбросом.

Источники

- NIST: исследовательский анализ (https://www.itl.nist.gov/div898/handbook/eda/eda.htm).
- IDS Skoltech: EDA (https://ids.skoltech.ru/seminars/seminar01/02b-eda.html).
- Python: статистические функции (https://docs.python.org/3/library/statistics.html).