

Современные методы аналитики и визуализации Основы визуализации данных Лекция 7

Кирилл Сысоев

Обо мне

5+ лет в Big Data

HSE University

Senior Data Engineer

OneFactor/UZUM Data

Hadoop, Spark, ClickHouse, Kafka, Docker

Python/Scala, SQL



t.me/KRSysoev

krsysoev@edu.hse.ru

Взаимодействие

Общение:

Мой telegram – личные вопросы/консультации/рекомендации

Лекции + ДЗ:

Telegram-чат «НИС Современные методы аналитики и визуализаций» – после лекций буду туда публиковать материалы лекций и описание ДЗ с дедлайном

Сдача ДЗ:

Почта – в установленный дедлайн буду ждать письмо с вложением

Введение в визуализацию данных

Визуализация данных – это мощный инструмент, который позволяет интерпретировать сложные массивы информации, находить закономерности и принимать осознанные решения. Однако её эффективность зависит от правильного выбора методов и инструментов, а также от грамотного представления информации.

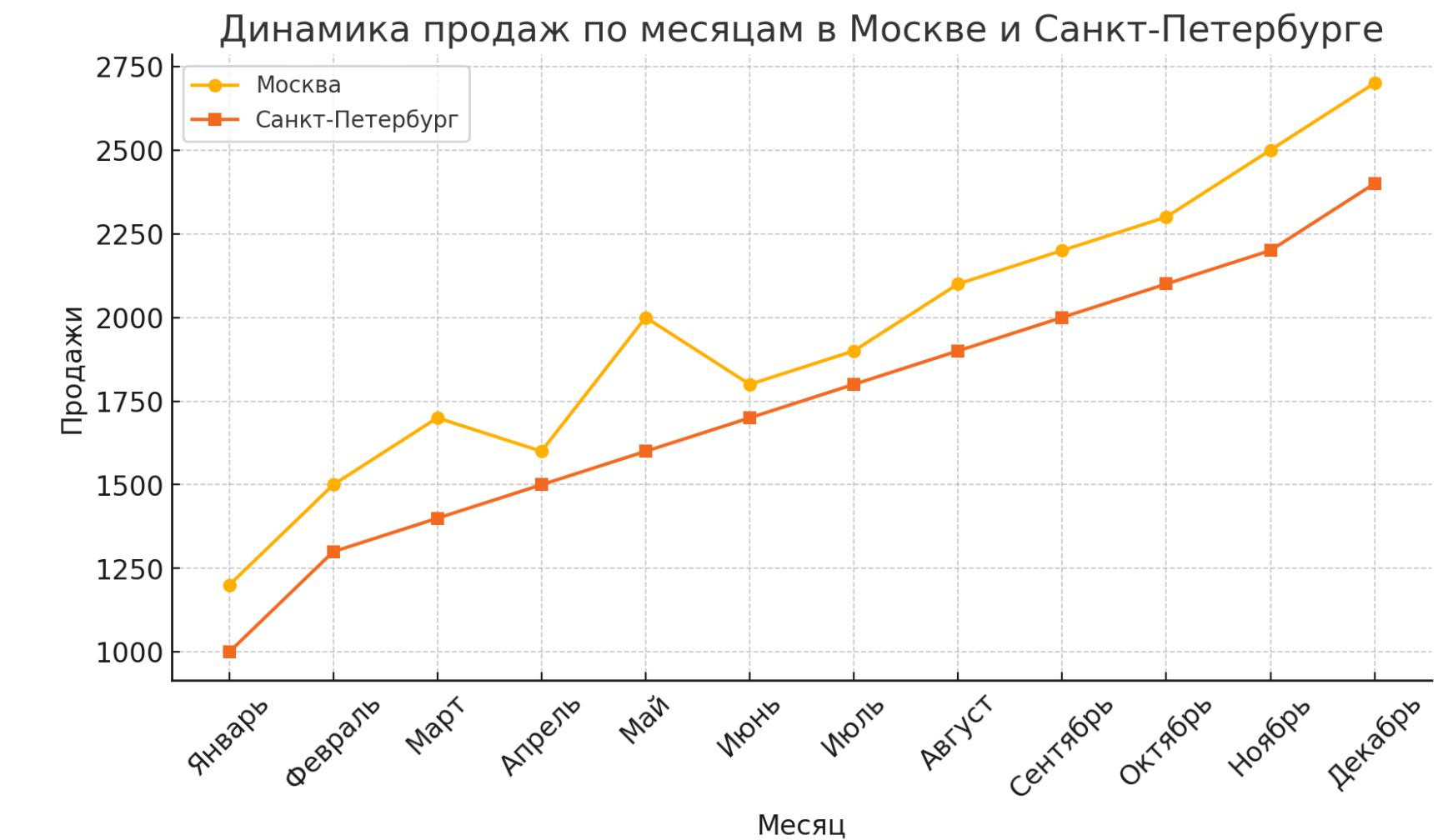


Важность визуализации в аналитике

Упрощение восприятия данных

Люди воспринимают визуальную информацию значительно быстрее, чем текст или таблицы чисел. Визуализация делает сложные данные интуитивно понятными.

Какой город показывает лучшие темпы роста?



Регион	Январь	Февраль	Март	Апрель	Май	Июнь	Июль	Август	Сентябрь	Октябрь	Ноябрь	Декабрь
Москва	1200	1500	1700	1600	2000	1800	1900	2100	2200	2300	2500	2700
СПб	1000	1300	1400	1500	1600	1700	1800	1900	2000	2100	2200	2400

Важность визуализации в аналитике

Выявление закономерностей и трендов

Графическое представление позволяет легко заметить тренды, корреляции и аномалии.

Пример: график тренда продаж может показать сезонность. Закономерные всплески, могут сигнализировать о сезонных распродажах или праздниках.



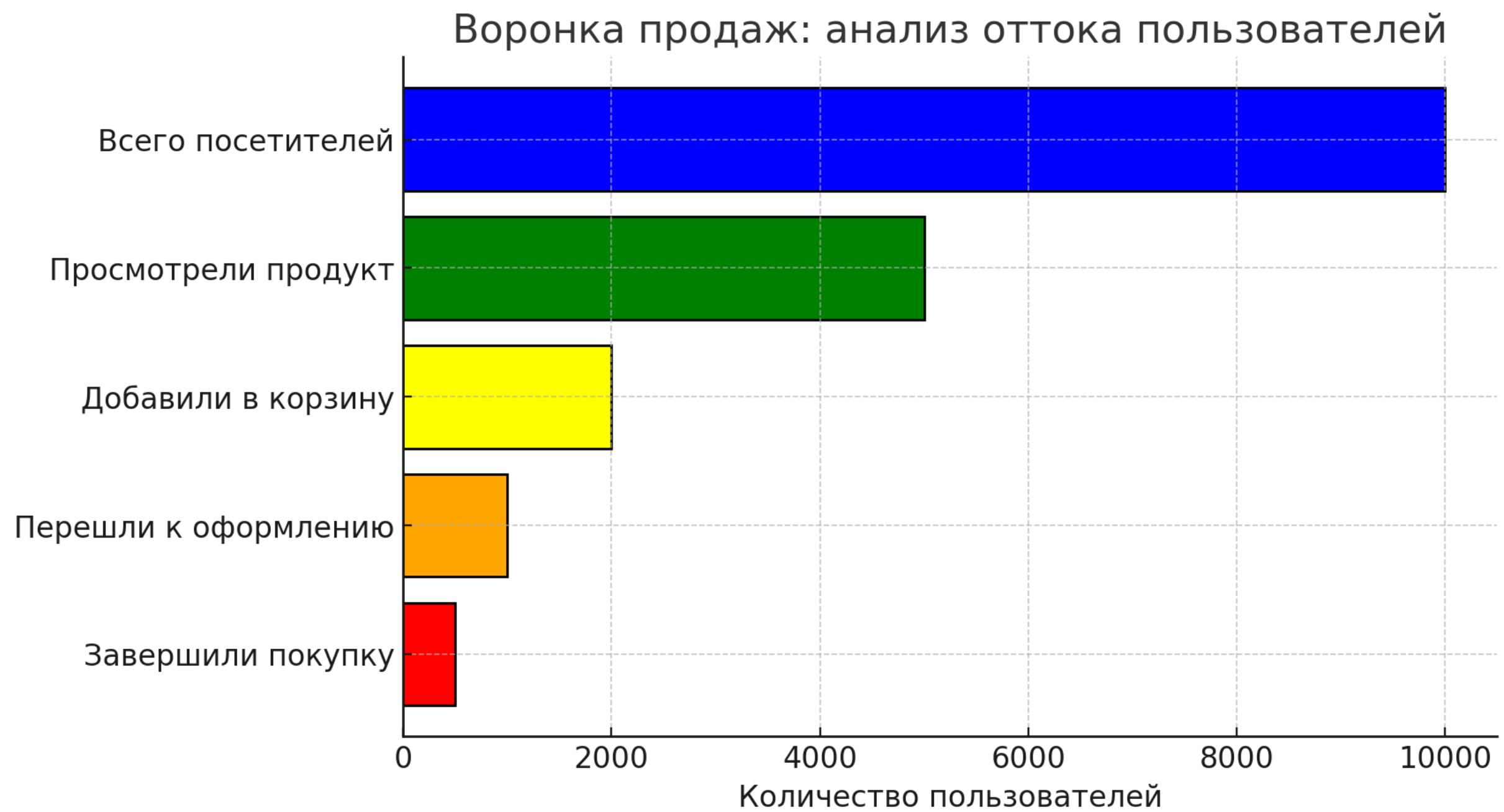
line chart

Важность визуализации в аналитике

Быстрое принятие решений

Менеджеры и руководители не будут читать длинные отчеты. Хорошая визуализация позволяет мгновенно оценить ситуацию и принять меры.

Пример: **визуализация воронки продаж по оттоку пользователей**. Видно, что наибольший отток происходит на этапе регистрации, что может сигнализировать о проблеме в пользовательском интерфейсе или сложности процесса оформления.



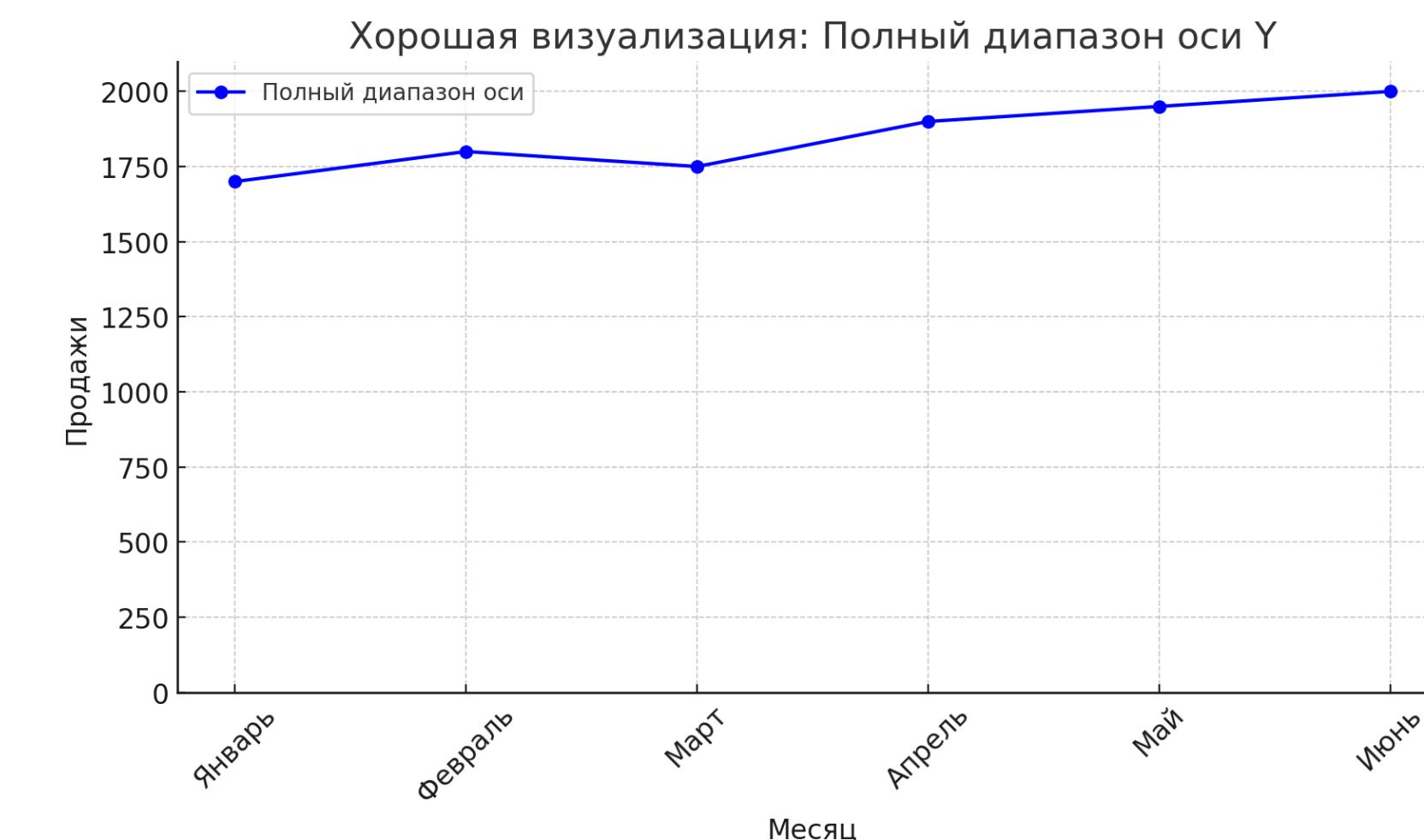
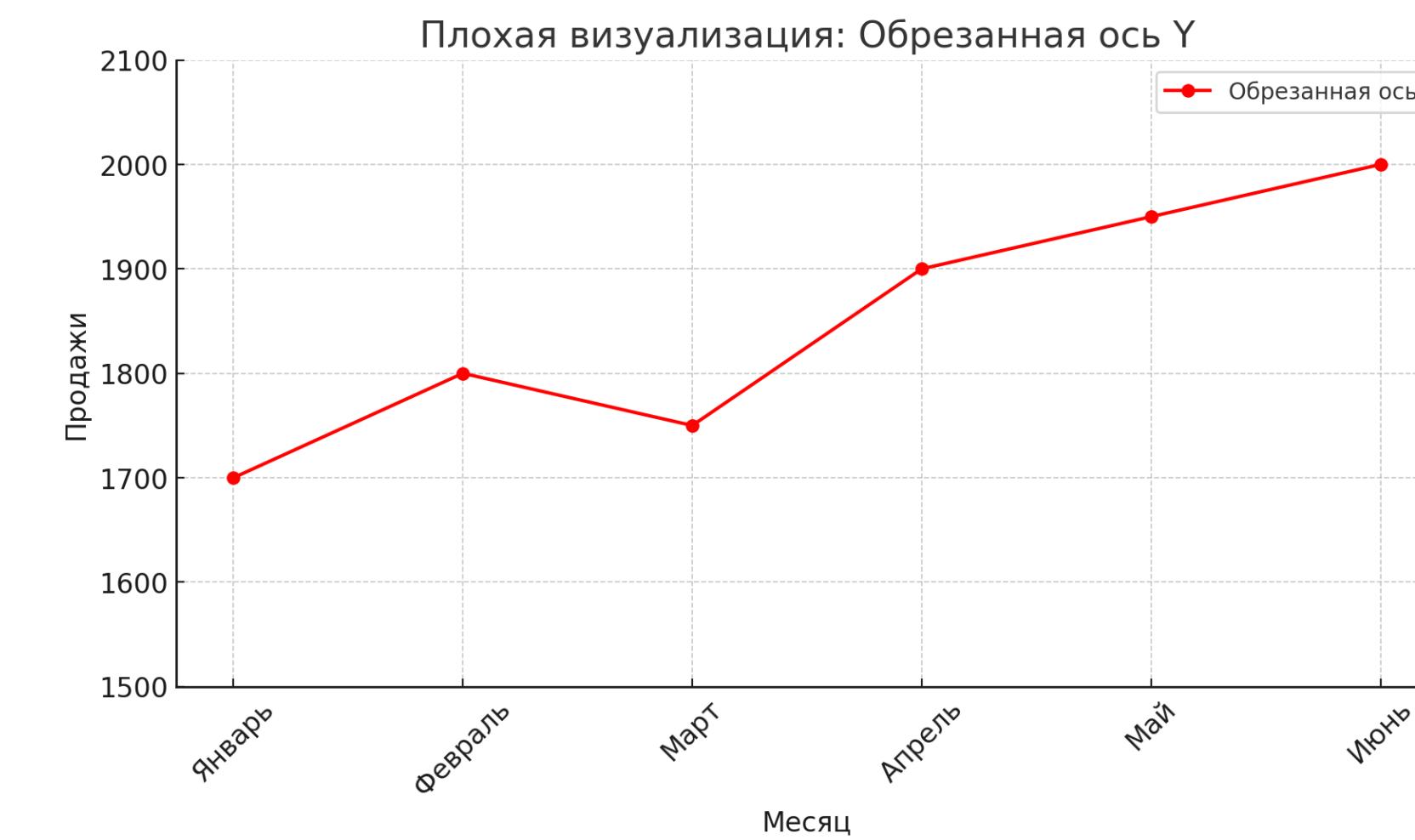
Как избежать искажений в данных?

Обрезка осей

Ошибка: Начало оси Y не с 0. Это создает иллюзию сильных различий.

Пример плохой визуализации: Диаграмма продаж, где ось Y начинается с 1500, показывает разницу между 1700 и 2000 как огромную.

Как исправить? Использовать полный диапазон, начиная с 0, чтобы не вводить в заблуждение.



Как избежать искажений в данных?

Допустимые случаи обрезки оси Y

1. Показ небольших изменений в больших числах

Если данные варьируются в узком диапазоне (например, от 98% до 99%), начинать ось с 0 может сделать график слишком плоским и неинформативным.

2. Финансовые данные

В финансовой аналитике часто используются обрезанные оси, чтобы показать динамику изменения цены акций, прибыли и других метрик. График акций, начинающийся с 0, может сделать даже значительные изменения (например, рост на 10%) визуально незаметными.

3. Научные данные

В научных исследованиях оси могут быть обрезаны для улучшения читаемости, особенно если данные анализируются в узком диапазоне.

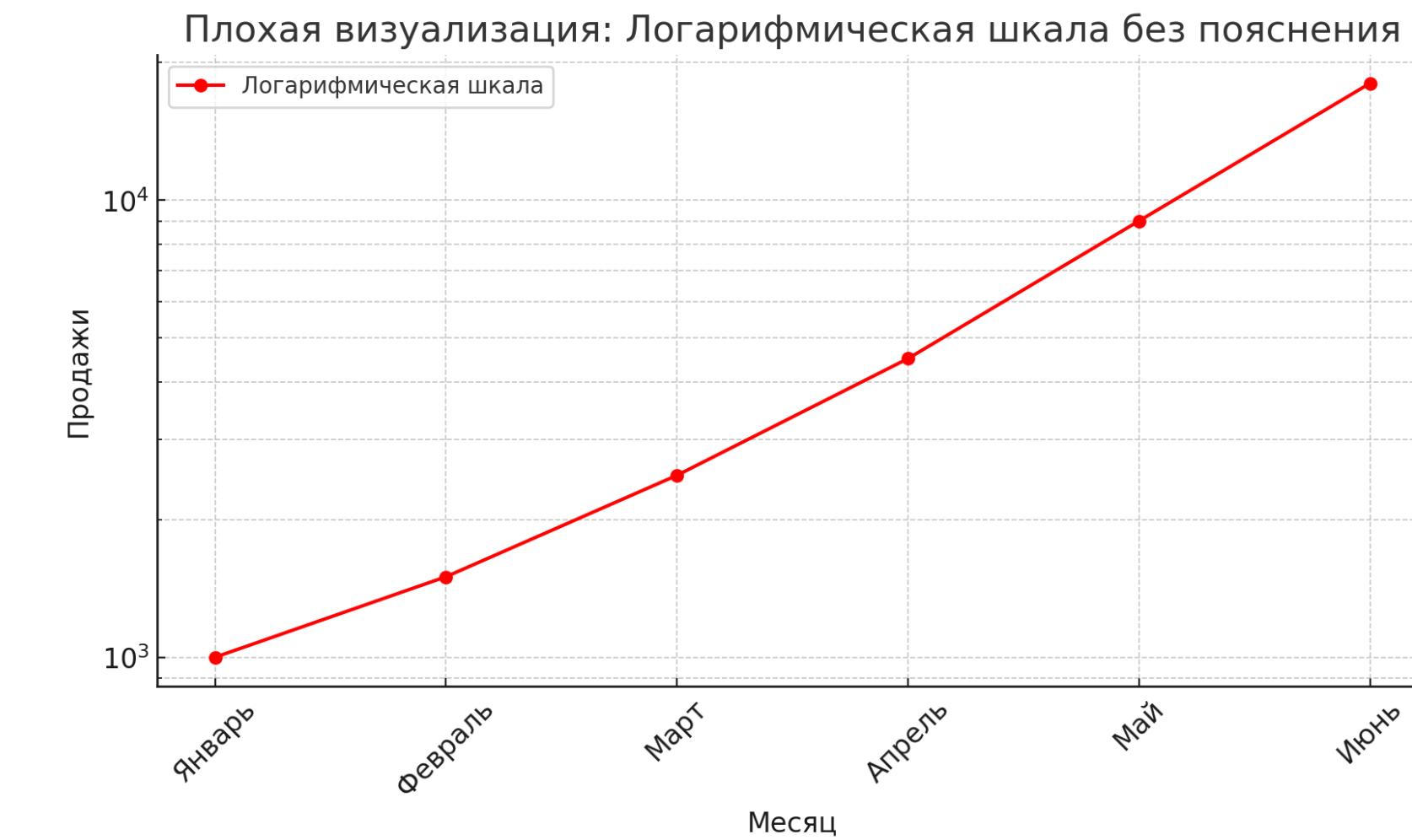
Как избежать искажений в данных?

Искажение масштаба

Ошибка: Использование логарифмической шкалы без пояснения.

Пример плохой визуализации: График показывает "плавный" рост продаж, но на самом деле там **логарифмическая шкала**, и реальный рост намного сильнее.

Как исправить? Объяснить использование логарифмов или выбрать линейный масштаб.



Как избежать искажений в данных?

Использование 3D-графиков

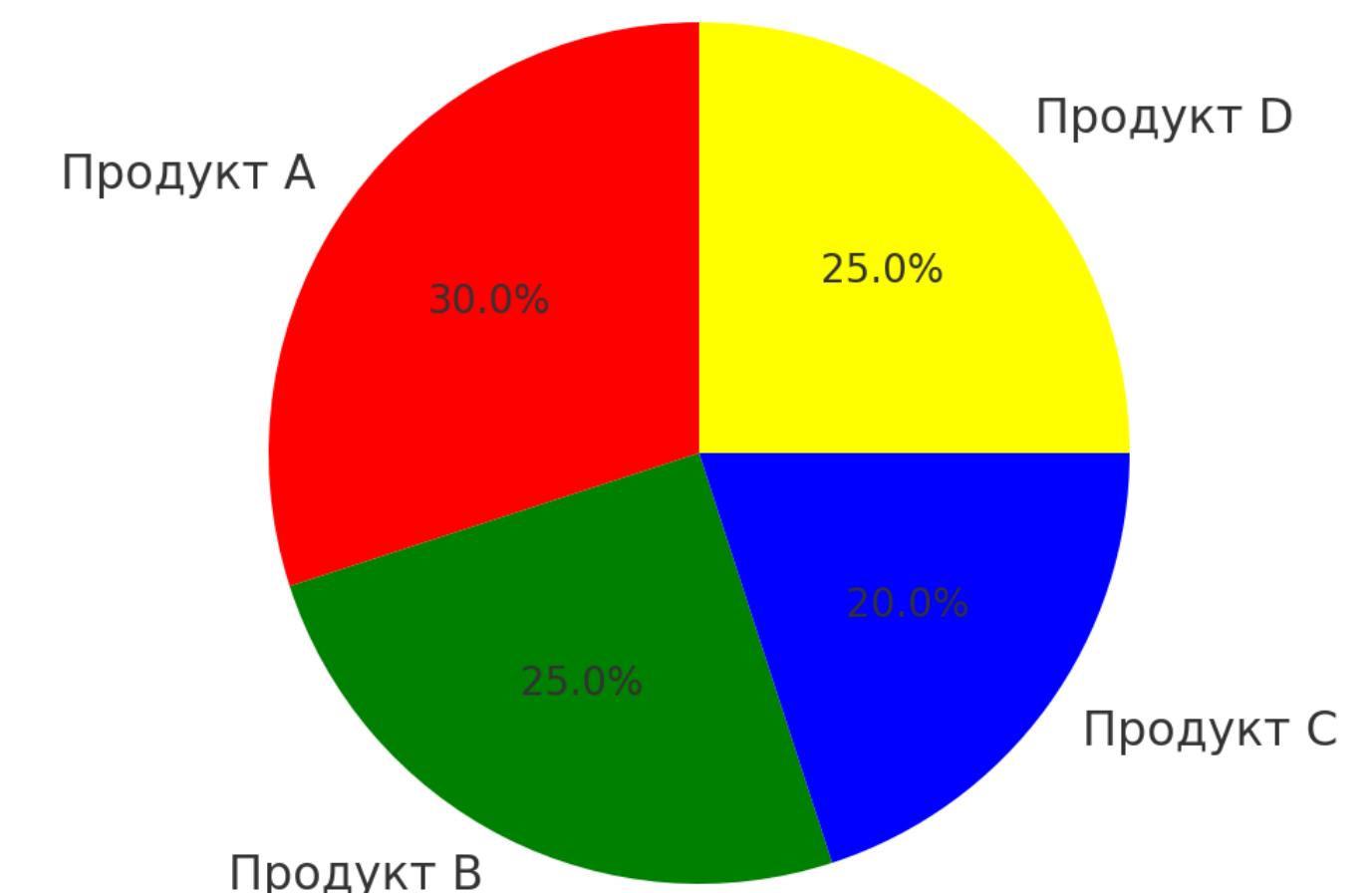
Ошибка: 3D-графики искажают пропорции и могут запутать зрителя.

Пример плохой визуализации: 3D-круговая диаграмма, где один сектор кажется больше только из-за перспективы.



Хорошая визуализация: 2D круговая диаграмма

Как исправить? Использовать плоские (2D) диаграммы.



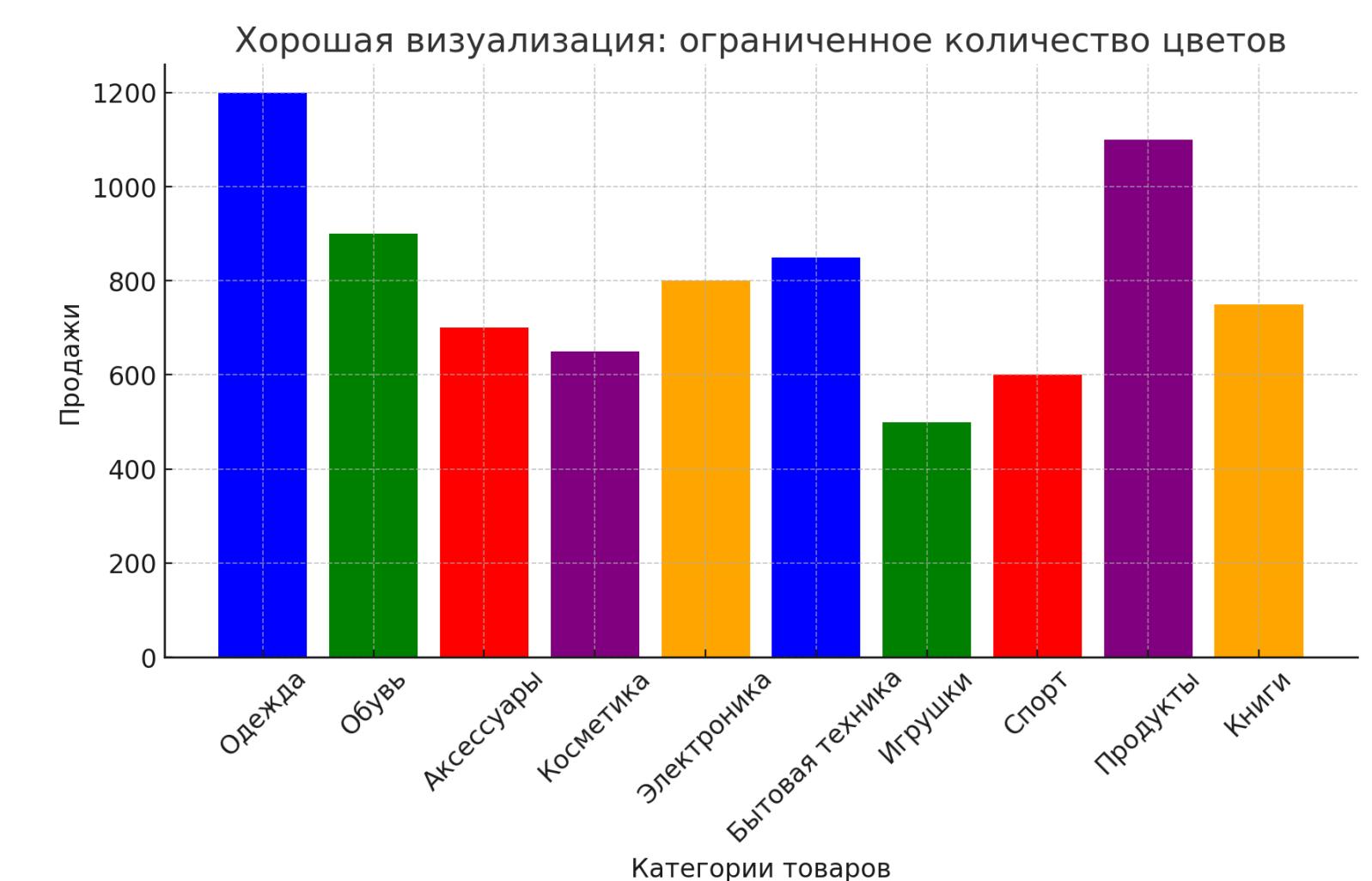
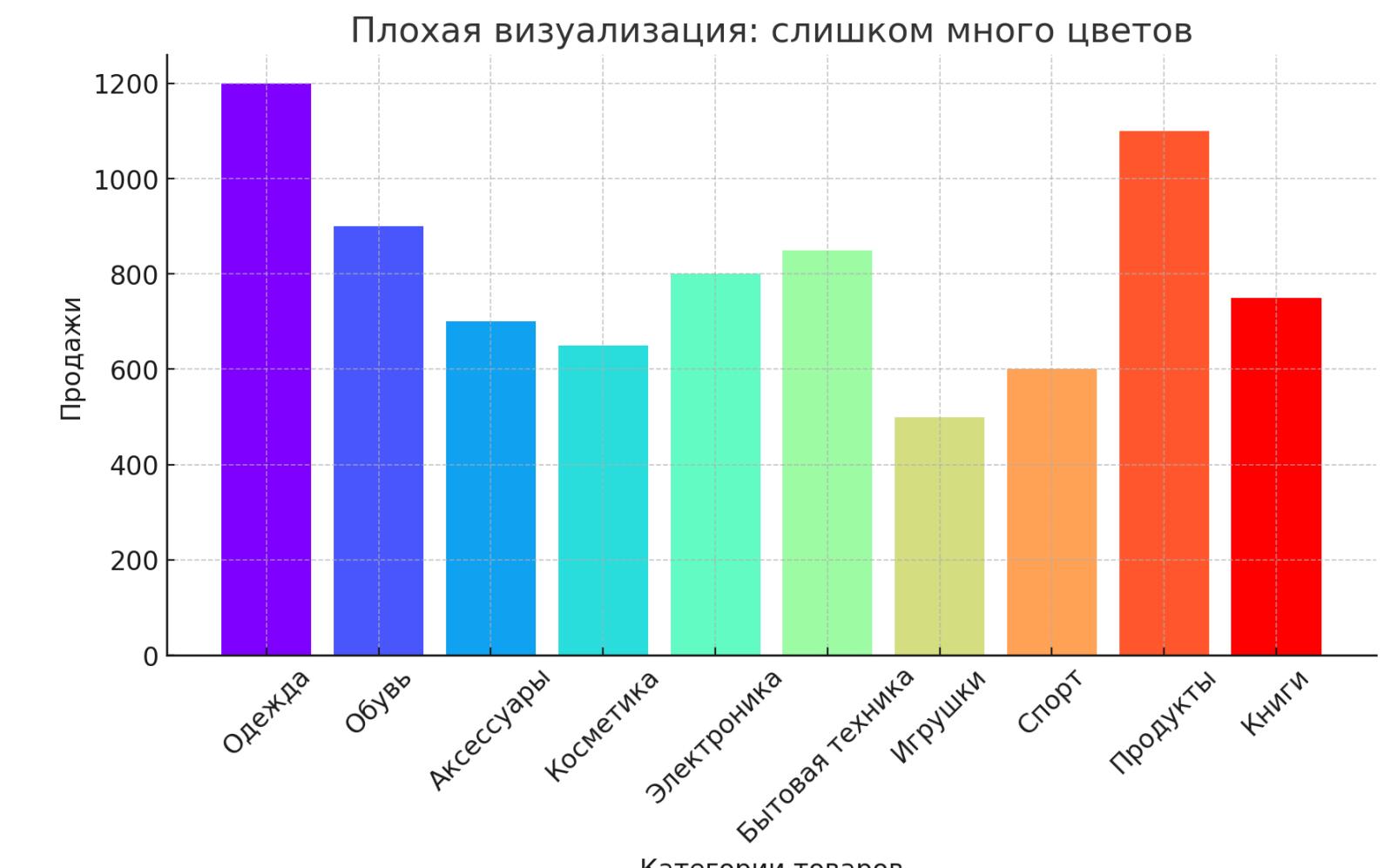
Как избежать искажений в данных?

Использование большого количества цветов

Ошибка: Слишком много цветов затрудняют восприятие.

Пример плохой визуализации: Гистограмма с 10 разными цветами для категорий – сложно анализировать.

Как исправить? Ограничить количество цветов (3-5) и использовать понятные оттенки.



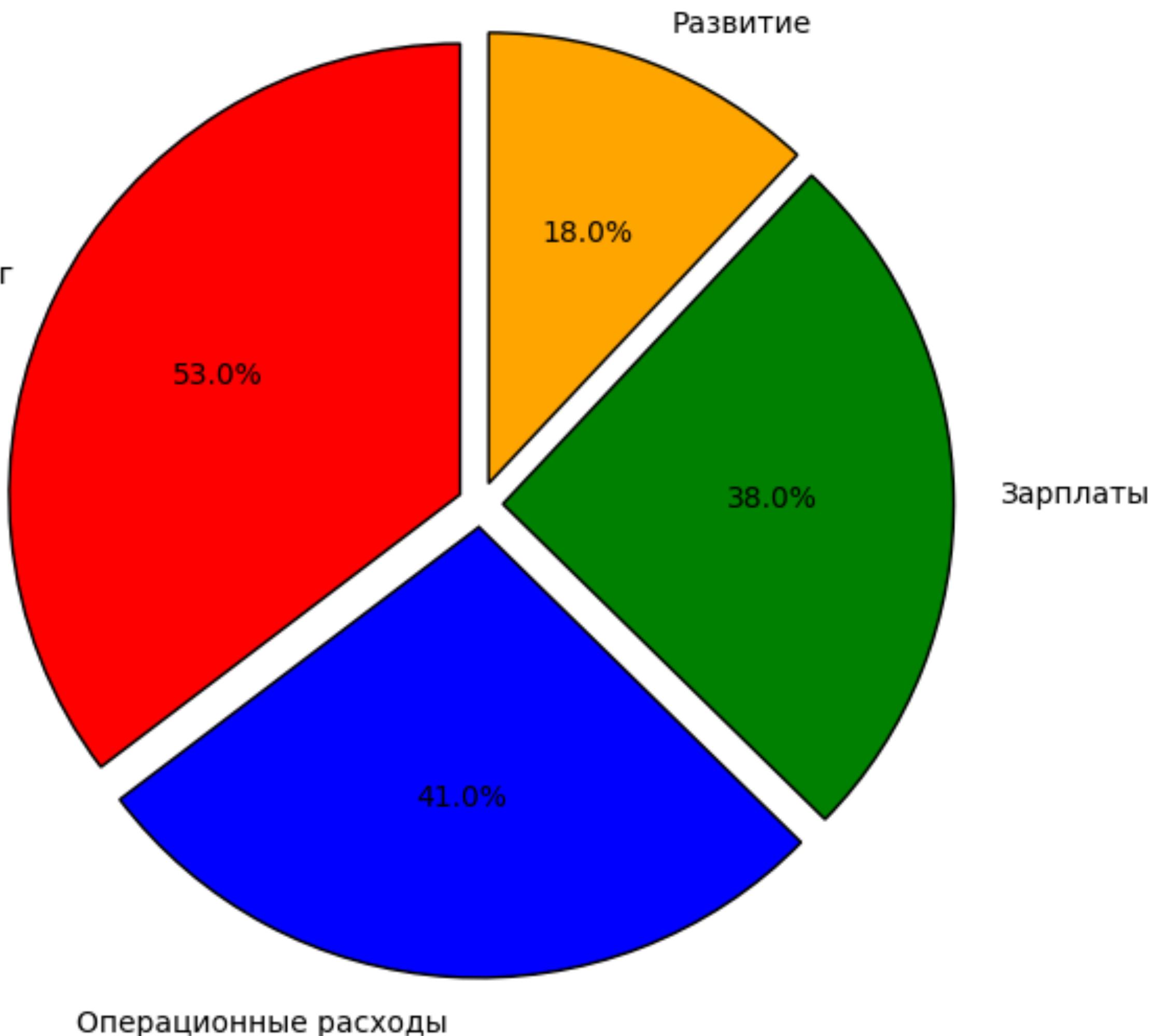
Кейс 1

Компания N представила отчет о распределении бюджета

Вопросы:

1. Что здесь не так?
2. Почему так случилось?
3. Как правильно отобразить данные?

Распределение бюджета компании

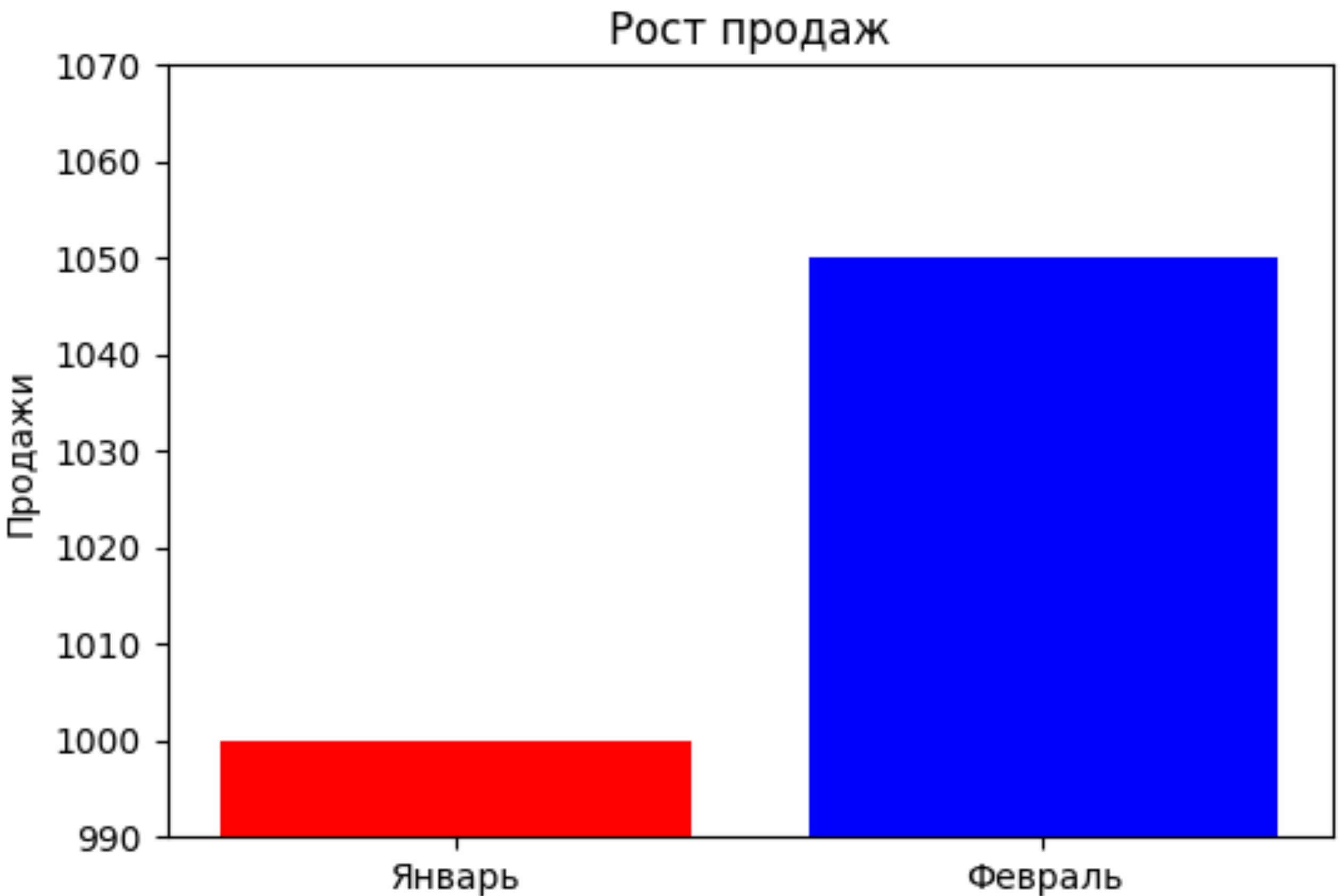


Кейс 2

График показывает рост продаж в магазине N за месяц

Вопросы:

1. В чем проблема с масштабом?
2. Почему такая визуализация может вводить в заблуждение?
3. Как можно сделать график более честным?



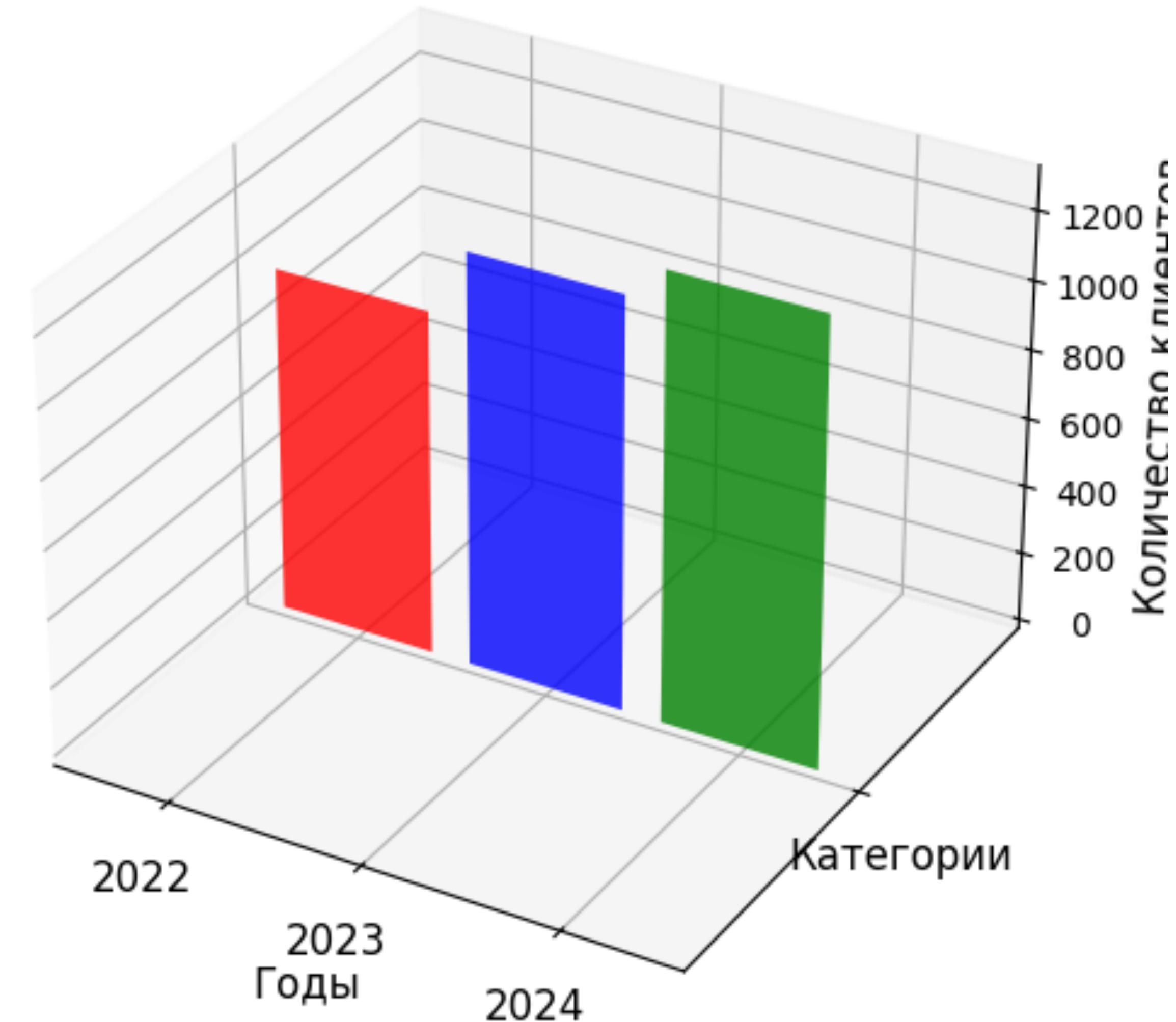
Кейс 3

Рост клиентов

Компания N представила данные о росте клиентов с помощью 3D-гистограммы

Вопросы:

1. В чем недостатки 3D-графиков?
2. Почему сложно интерпретировать такие данные?
3. Какие альтернативы можно использовать?



Основные принципы визуализации данных

Грамотная визуализация **должна помогать анализу данных**, а не усложнять его.

Визуальные элементы должны быть интуитивно понятны, минимизировать когнитивную нагрузку и передавать смысл без искажений.



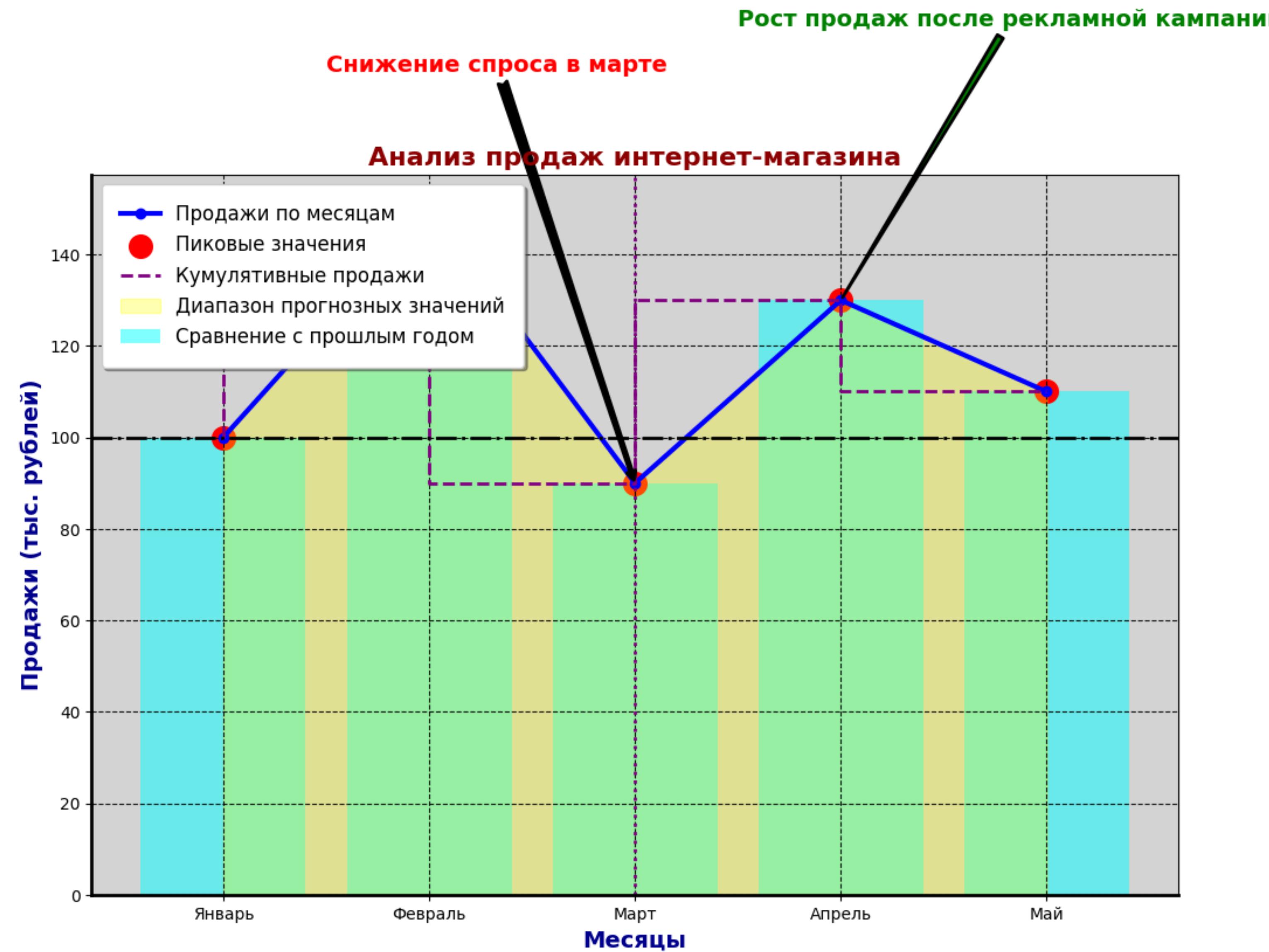
Как минимизировать когнитивную нагрузку?

Когнитивная нагрузка — это **объем информации**, который человек должен обработать, чтобы понять график. Чем больше лишних деталей, тем сложнее воспринять данные.

Ошибки, создающие когнитивную нагрузку:

1. Слишком много информации на одном графике
2. Использование сложных 3D-эффектов
3. Плохая читаемость из-за мелкого шрифта
4. Непонятные обозначения и сокращения

Как минимизировать когнитивную нагрузку?



Как уменьшить когнитивную нагрузку?

1. Убирать лишние элементы (data-ink ratio, концепция Эдварда Тафти)
 1. Убирайте ненужные линии, границы, сетки, если они не несут информации.
 2. Страйтесь минимизировать использование теней и объемных эффектов.
2. Использовать понятные подписи вместо легенд
 1. Подписывайте линии и категории прямо на графике, вместо того чтобы заставлять читателя искать соответствия.
3. Не перегружать визуализацию большим количеством категорий
 1. Например, если у вас есть 15 категорий, лучше сгруппировать их или использовать столбчатую диаграмму вместо круговой.
4. Выбирать правильный тип графика
 1. Для временных рядов лучше подходят линейные графики, а не гистограммы.
 2. Для сравнения частей целого удобнее стекированные столбчатые диаграммы, а не 3D-круговые.

Что важнее: красота или информативность?

При создании графиков важно находить **баланс** между эстетикой и понятностью данных.

Ошибки при перегибе в сторону красоты:

1. Слишком яркие и пестрые цвета отвлекают от смысла
2. Использование декоративных элементов, которые не несут информации
3. Графики с 3D-эффектами и сложными текстурами

Литература для доп погружения в тему

Спасибо за внимание!