

# **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](https://t.me/KRSysoev)

[krsysoev@edu.hse.ru](mailto: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.

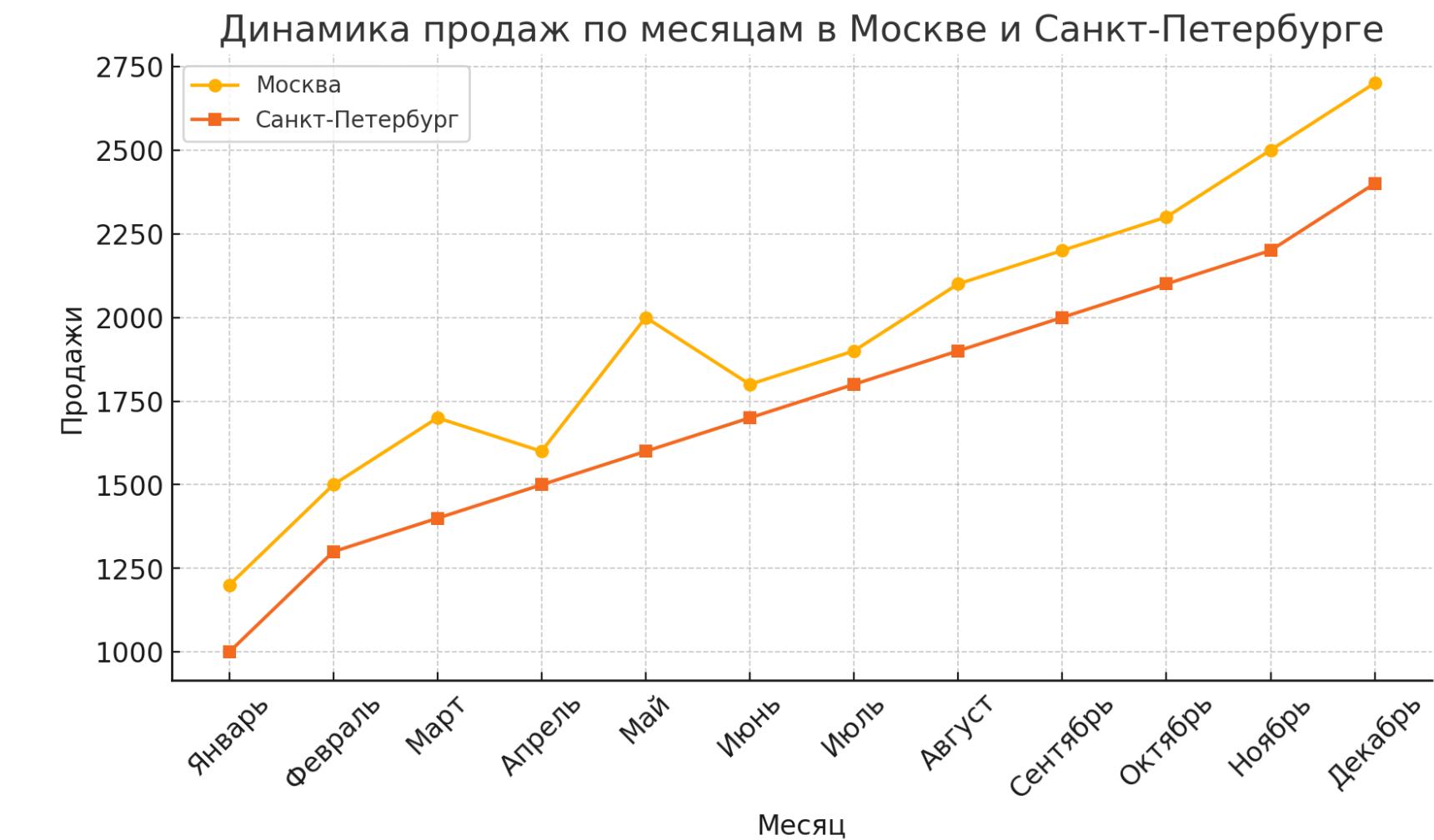


# Важность визуализации в аналитике

## 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.



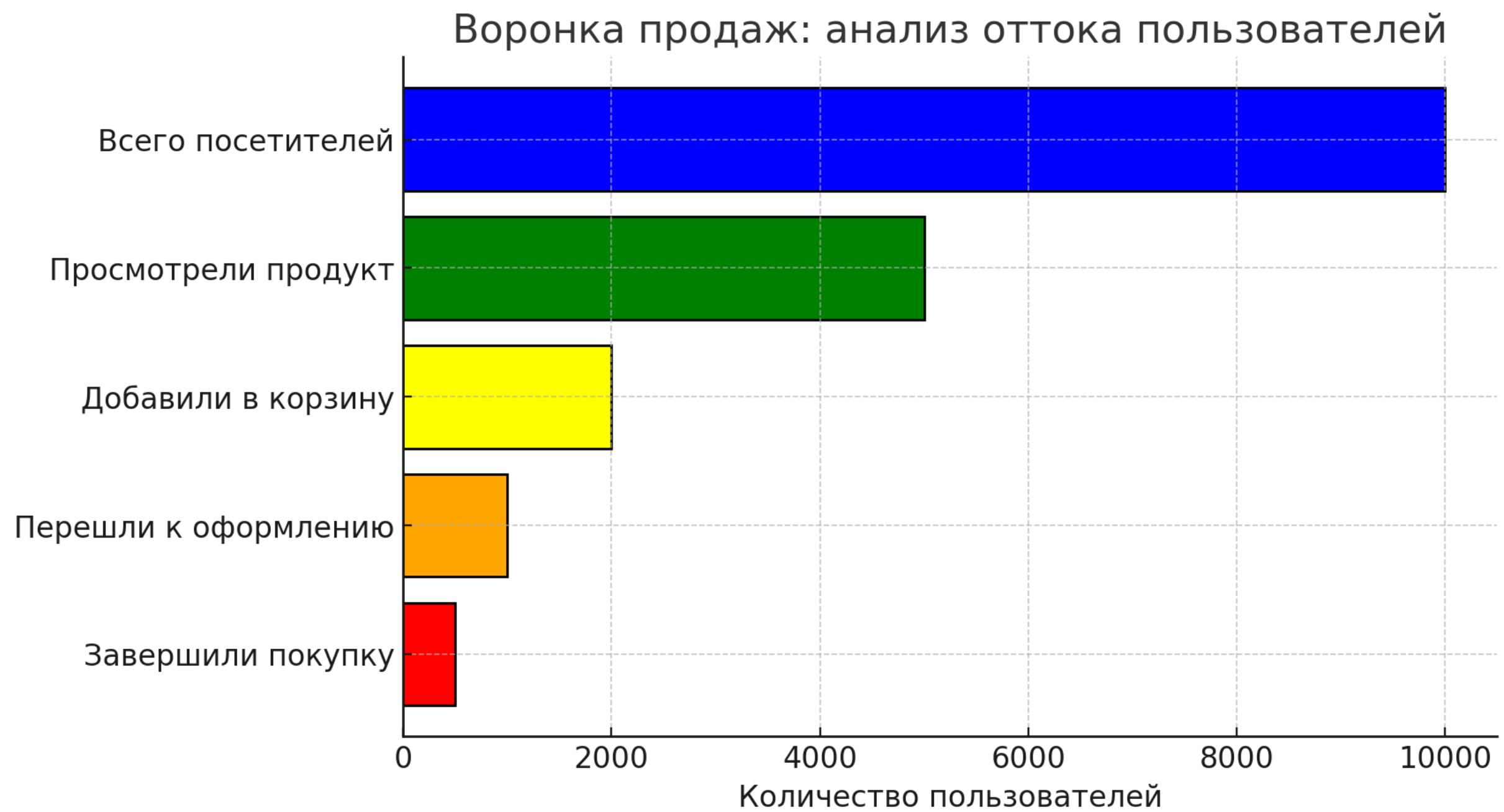
line chart

# 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.



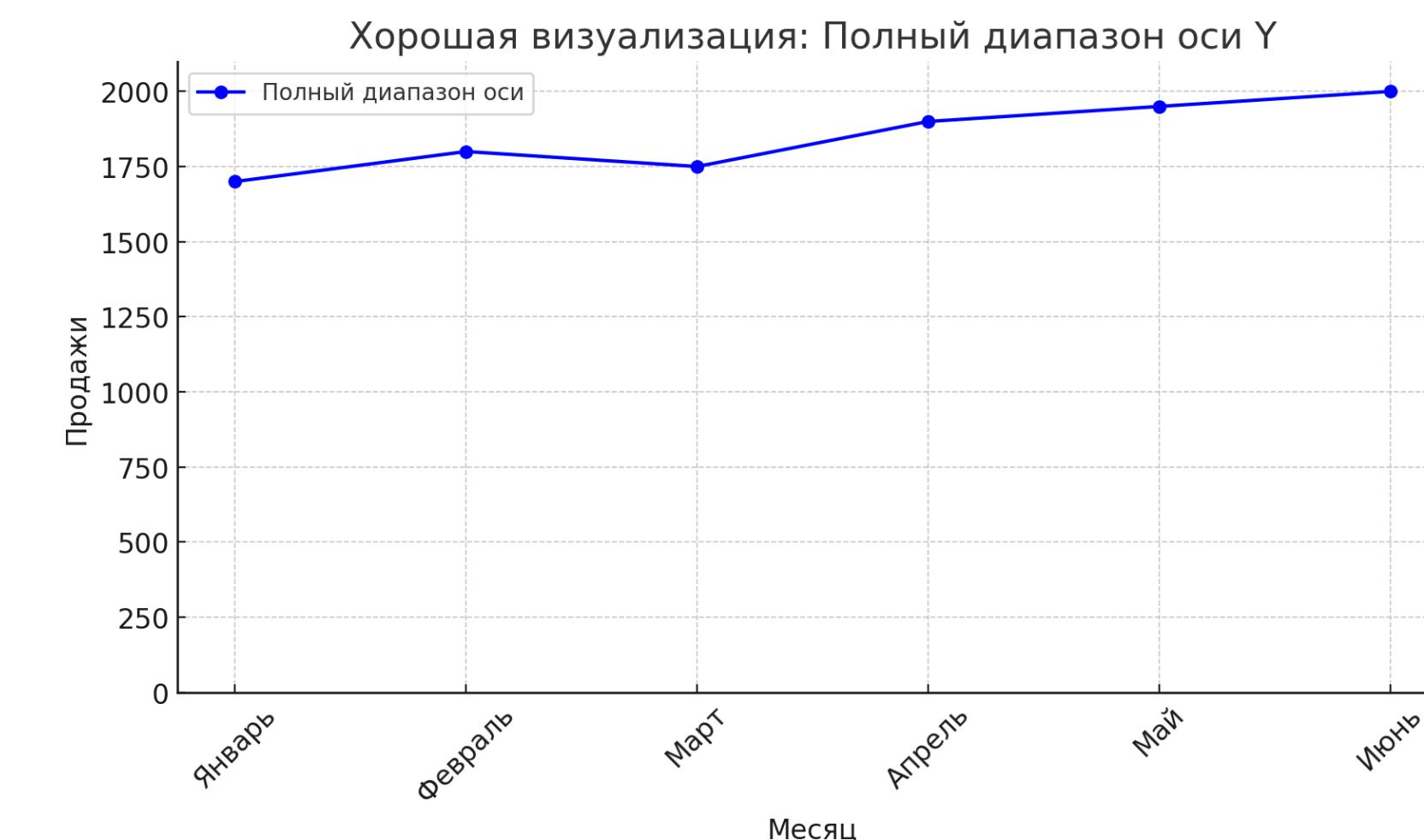
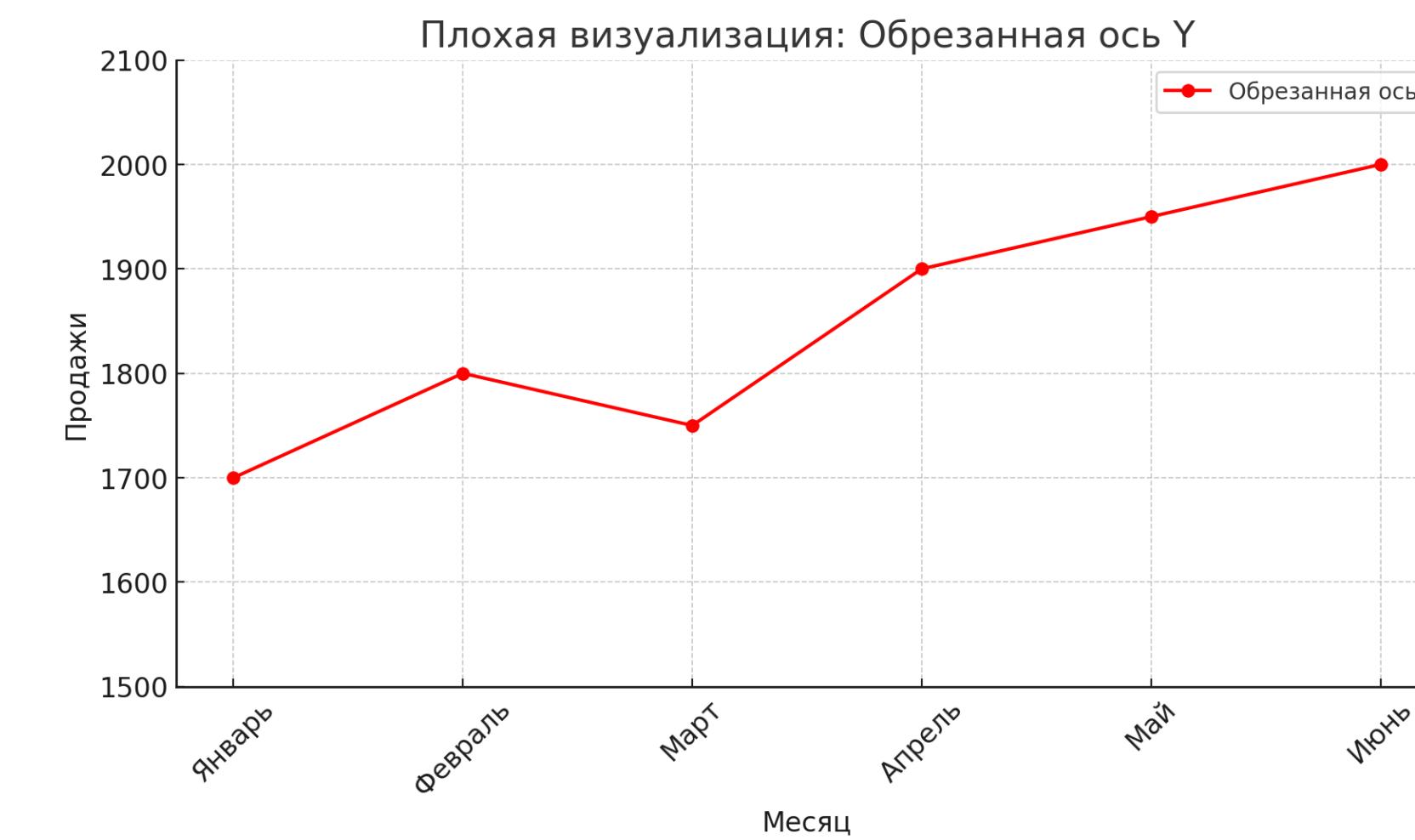
# Как избежать искажений в данных?

## 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.

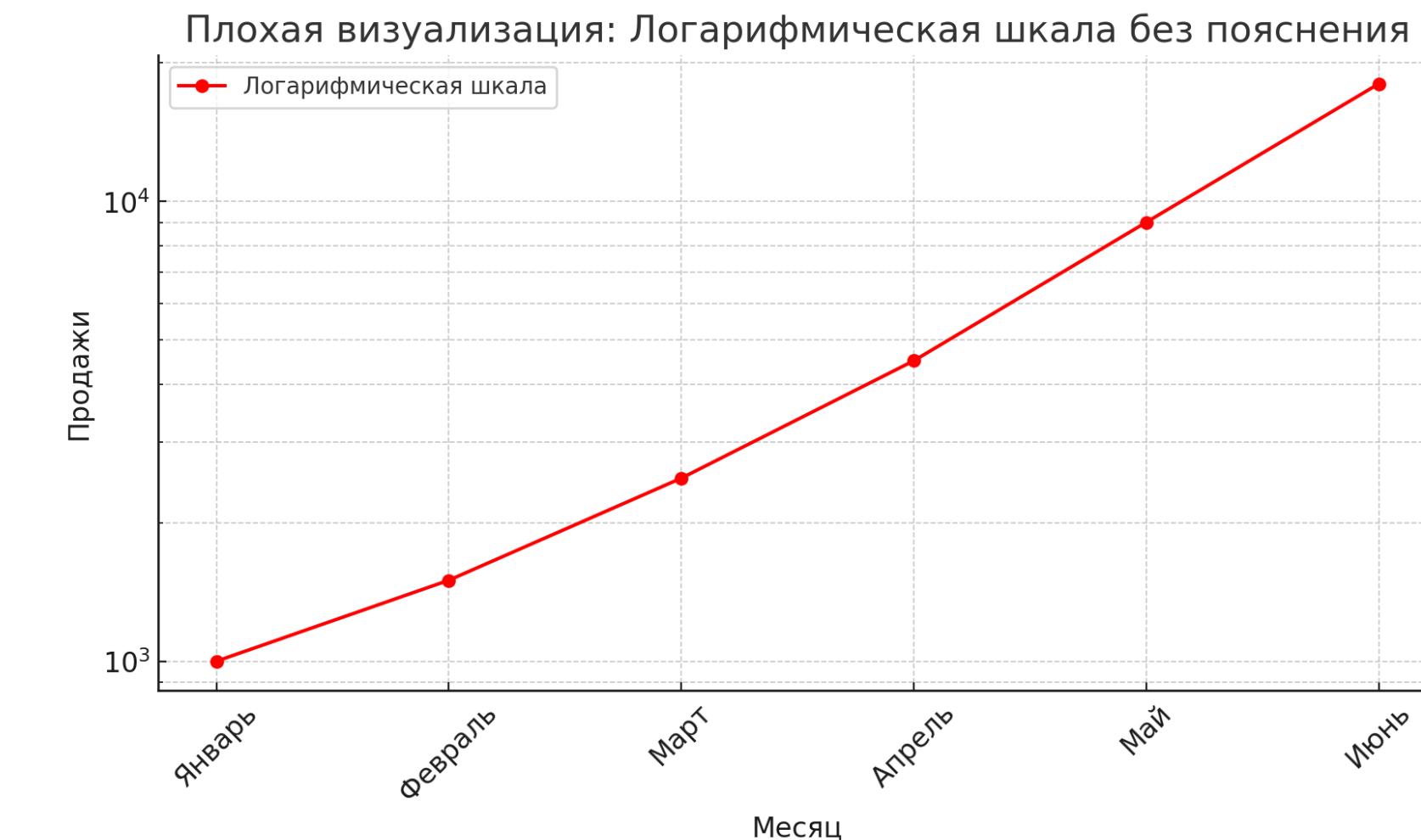
# Как избежать искажений в данных?

## 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

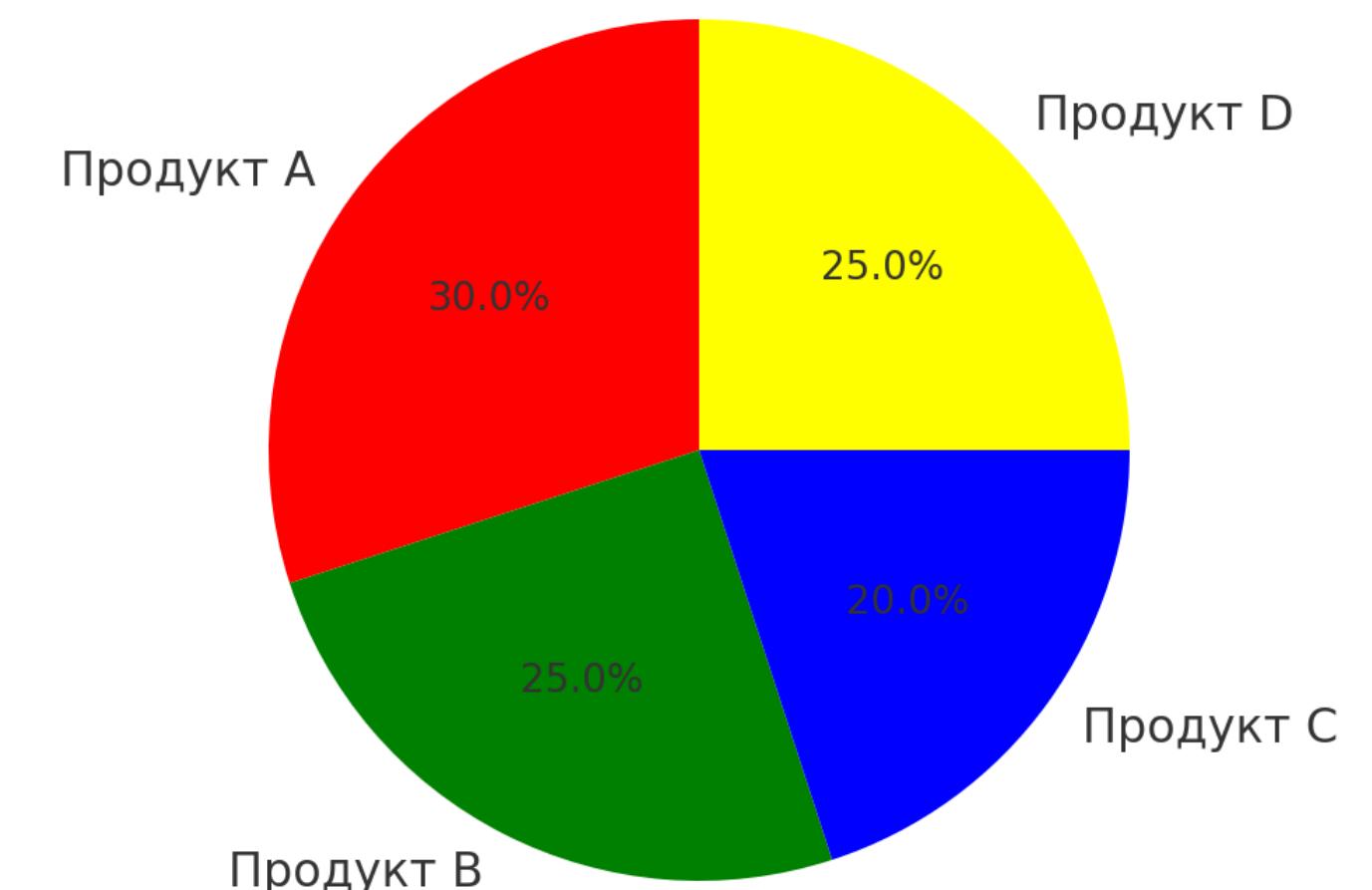
**Ошибка:** 3D-графики искажают пропорции и могут запутать зрителя.

**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.



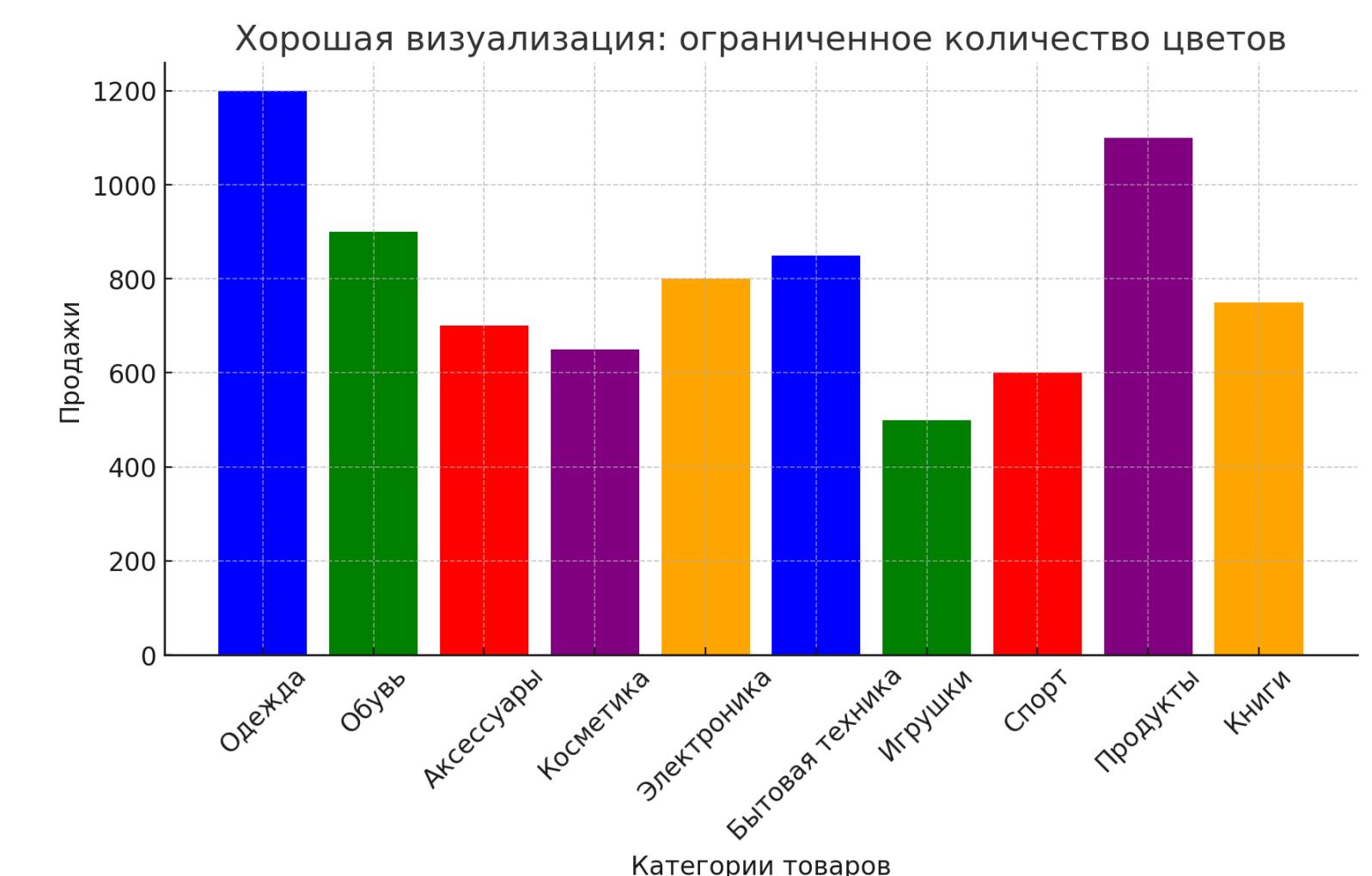
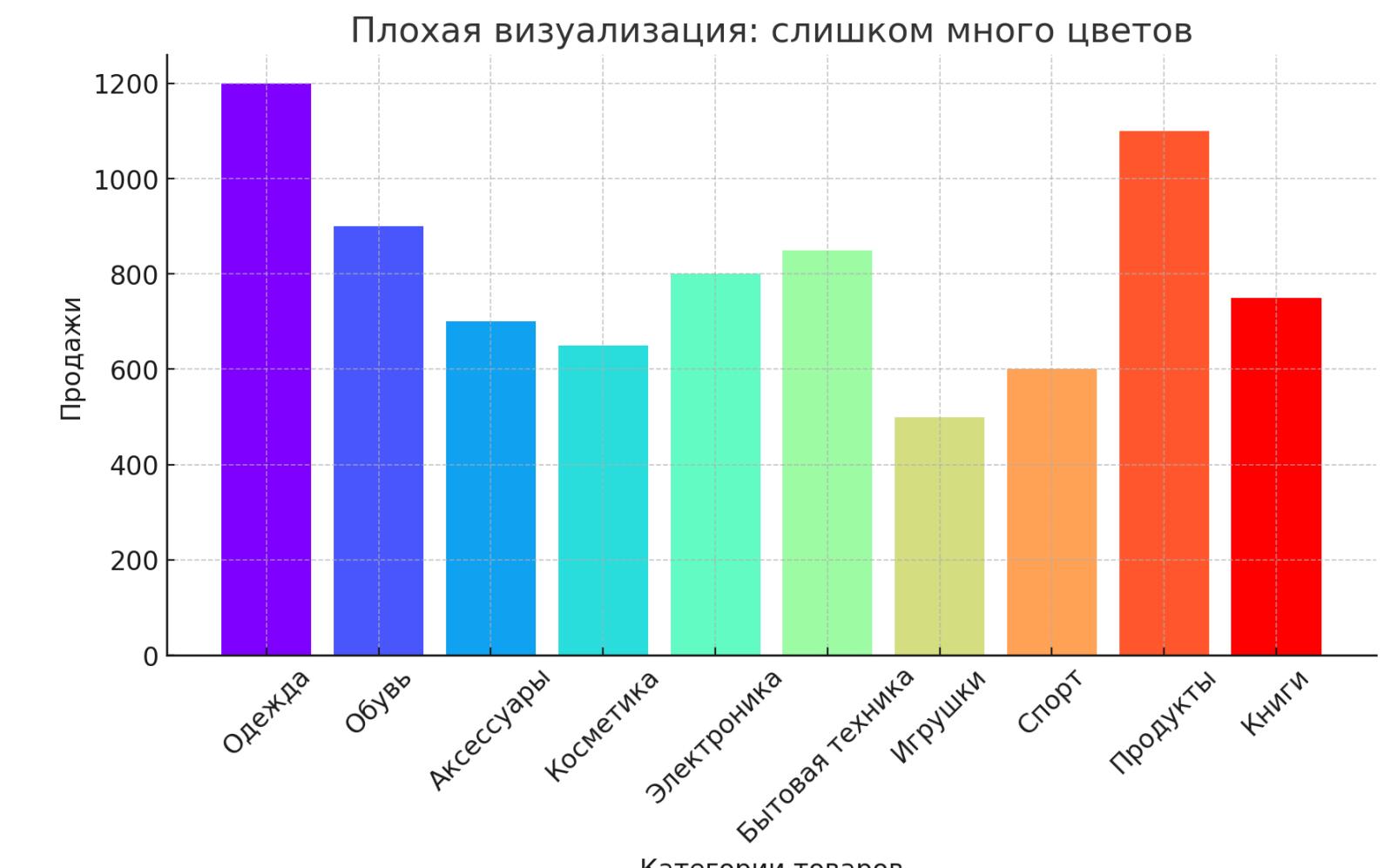
# Как избежать искажений в данных?

## 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.



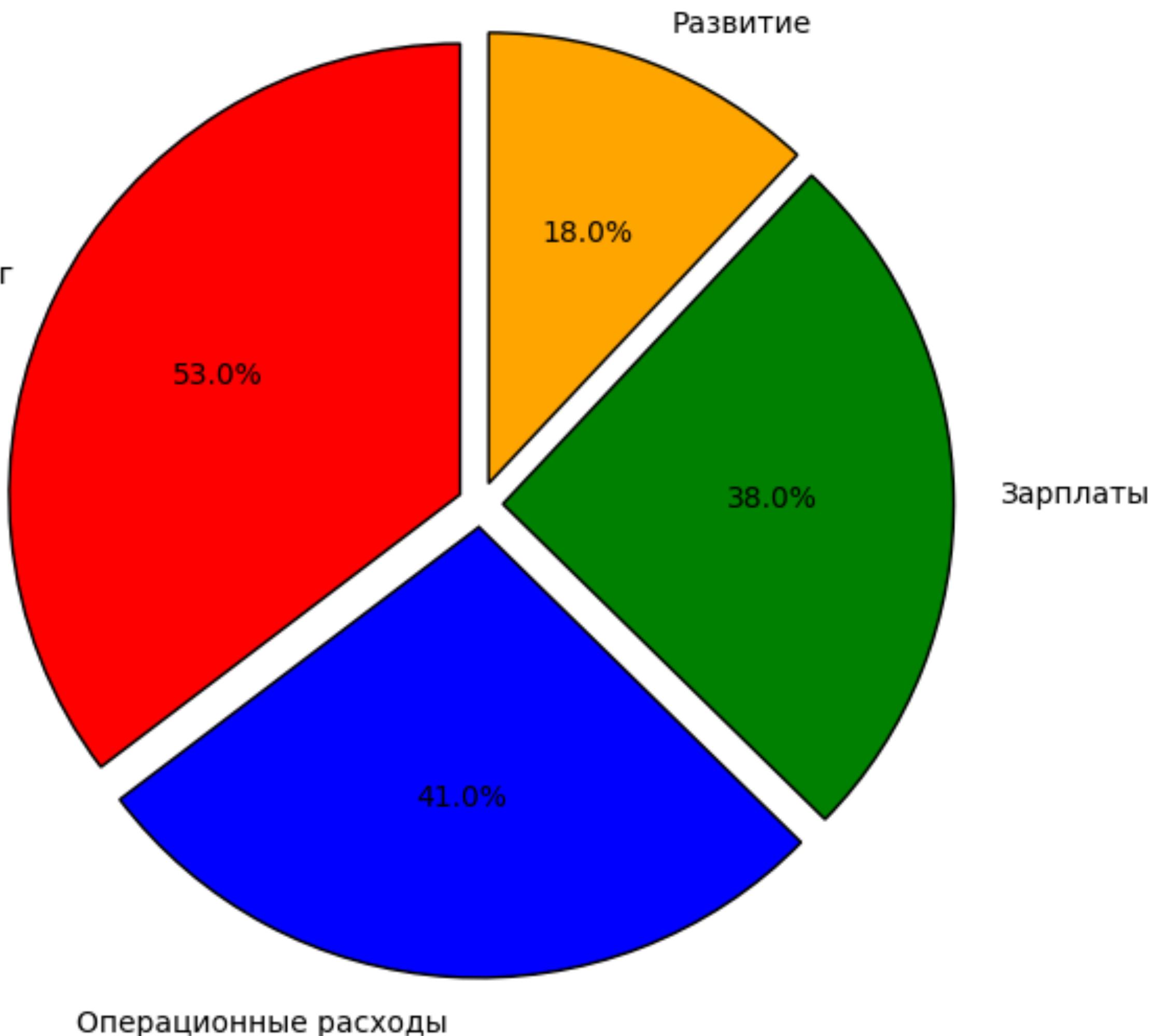
# Кейс 1

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

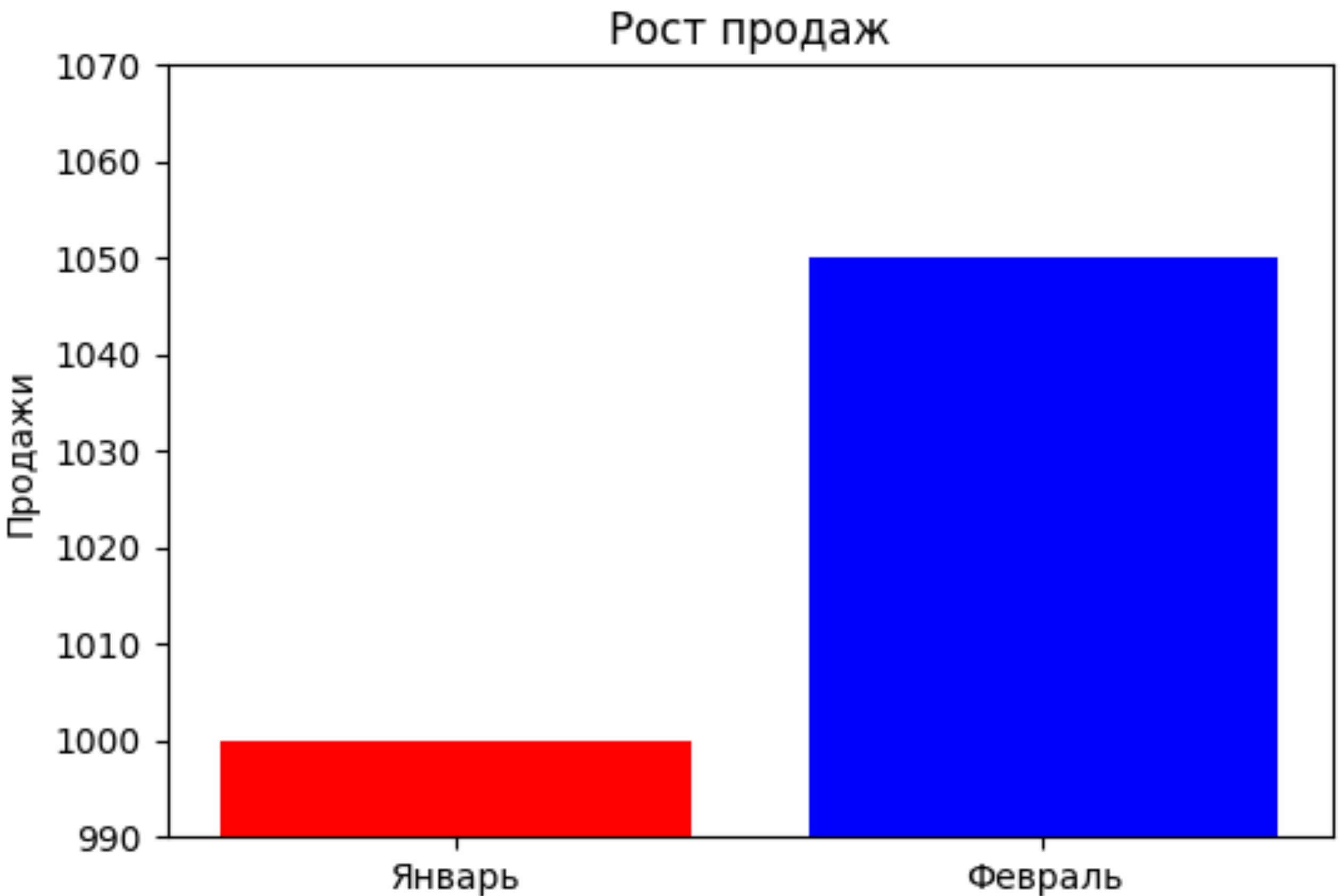


# Кейс 2

График показывает рост продаж в магазине N за месяц

Вопросы:

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



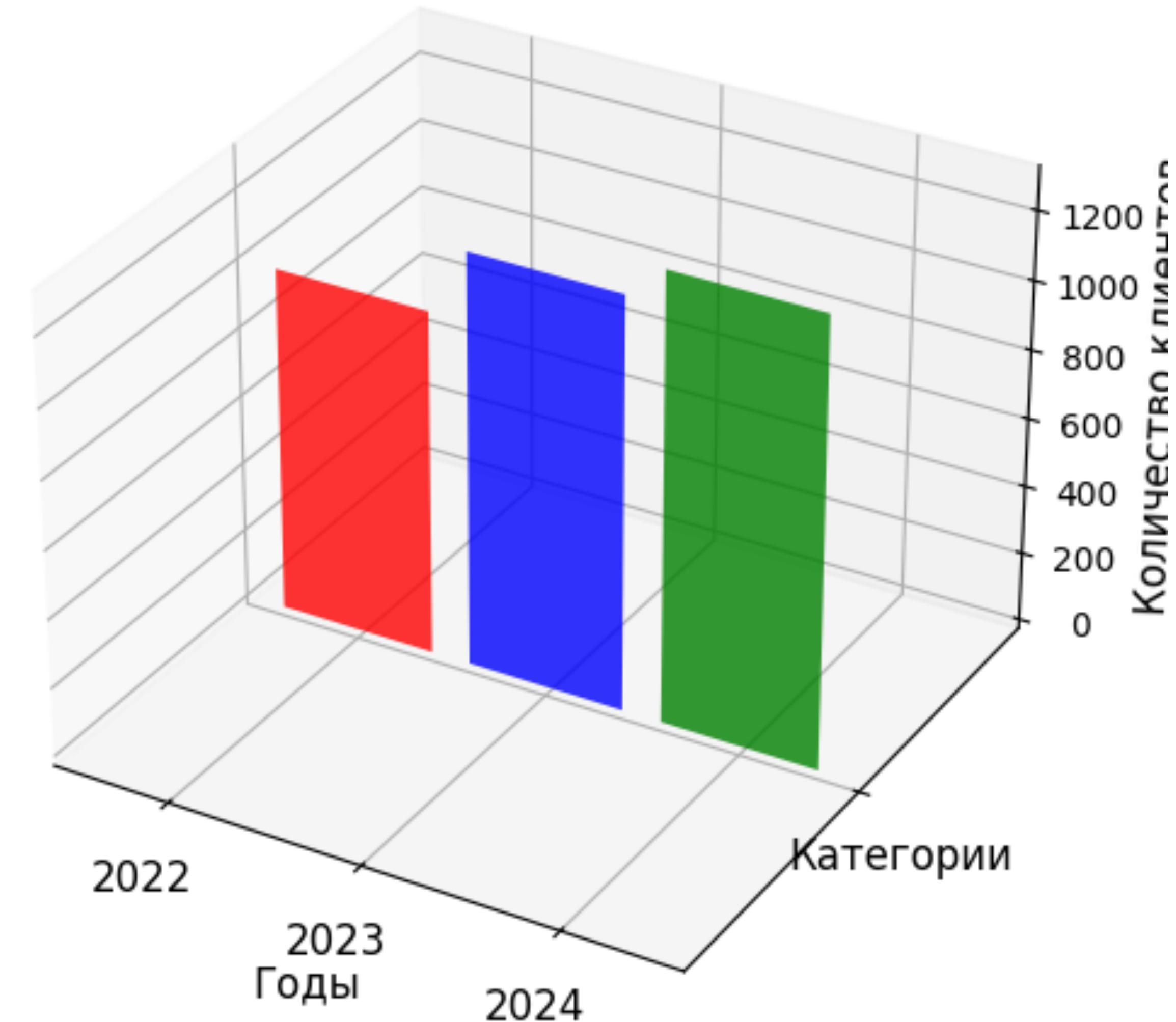
# Кейс 3

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.



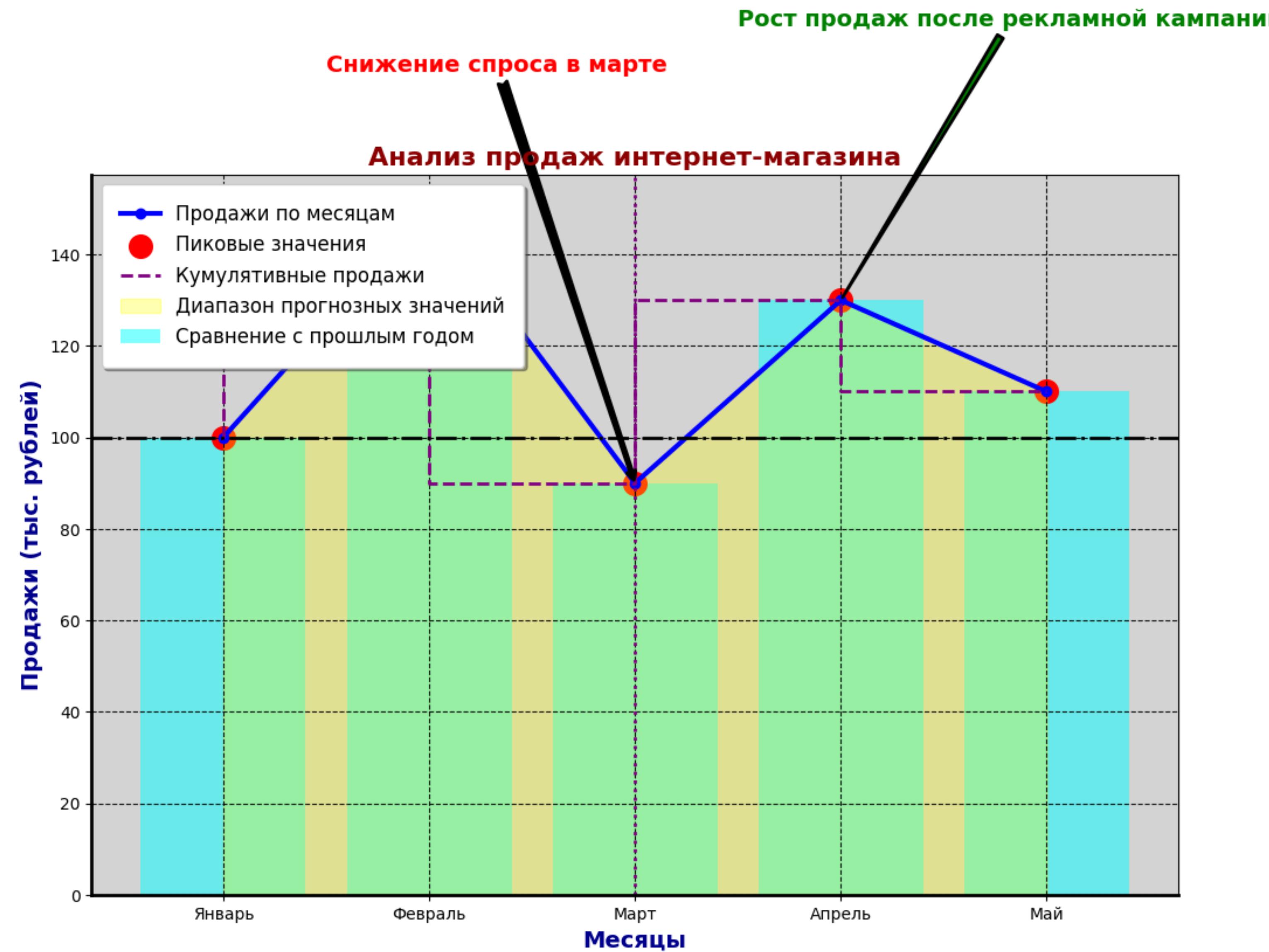
# Как минимизировать когнитивную нагрузку?

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 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

# **Как сделать визуализацию одновременно красивой и информативной?**

## **1. Фокус на данных, а не на оформлении**

1. Используйте простые, чистые стили.
2. Избегайте лишних украшений, теней и градиентов.

## **2. Выдерживать баланс между цветами и читаемостью**

1. Не перегружайте график слишком яркими цветами.
2. Используйте цвет только для выделения ключевых элементов.

## **3. Использовать правильные масштабы**

1. Не обрезайте оси, если это искажает данные.
2. Поддерживайте правильные пропорции, чтобы сохранить реальную картину.

# Работа с цветовыми схемами и форматами данных

Цвет — мощный инструмент, но его нужно использовать с умом.

## Ошибки при выборе цвета:

1. Слишком много цветов — сложно анализировать данные
2. Несоответствие цвета и смысла (например, красный для роста продаж)

# **Как правильно использовать цвета?**

## **1. Использовать цвет для выделения, а не для украшения**

1. Один акцентный цвет (например, синий) на фоне нейтральных цветов поможет выделить ключевые данные.

## **2. Соблюдать логику цвета**

1. Красный = снижение, убытки
2. Зеленый = рост, положительная динамика
3. Желтый = предупреждение

## **3. Использовать цветовые палитры**

1. Paletton, ColorBrewer помогут выбрать безопасные цвета.

# Типы графиков и их применение

Выбор правильного типа графика – один из ключевых этапов визуализации данных. Разные типы графиков **подходят для разных типов данных:** числовых, категориальных, временных или географических.

.



# Числовые данные

**Числовые (количественные) данные** – это данные, представленные числами, например, возраст, доход, количество клиентов.

**Виды диаграмм:**

1. Гистограммы
2. Boxplot
3. Scatter plot

# Числовые данные

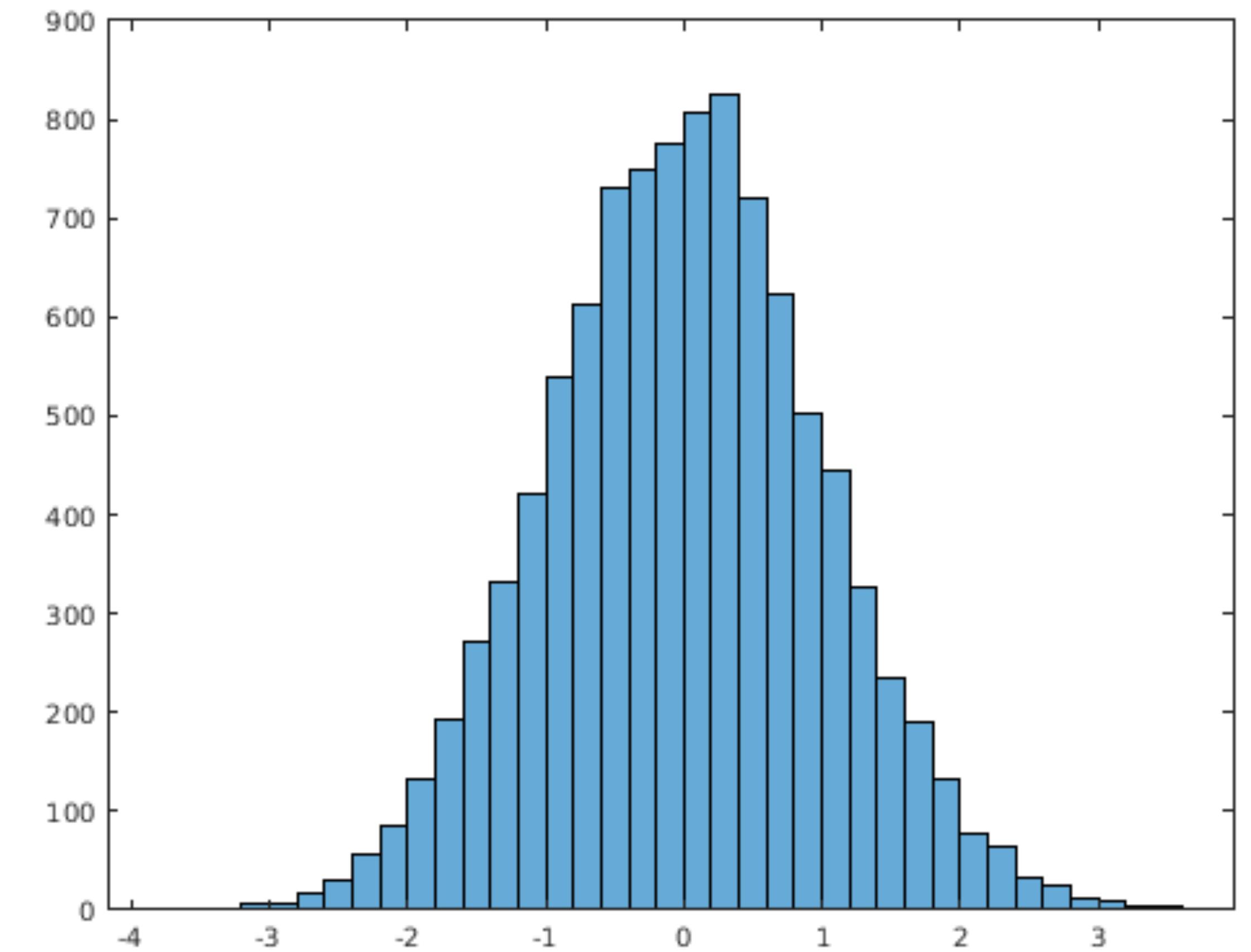
## Гистограммы (Histogram)

### Когда использовать?

- Для отображения распределения числовых данных.
- Когда важно понять, есть ли выбросы или нормальное распределение.

### Как интерпретировать?

- Если гистограмма симметрична, данные имеют нормальное распределение.
- Длинный "хвост" справа или слева указывает на асимметричность распределения.



# Числовые данные

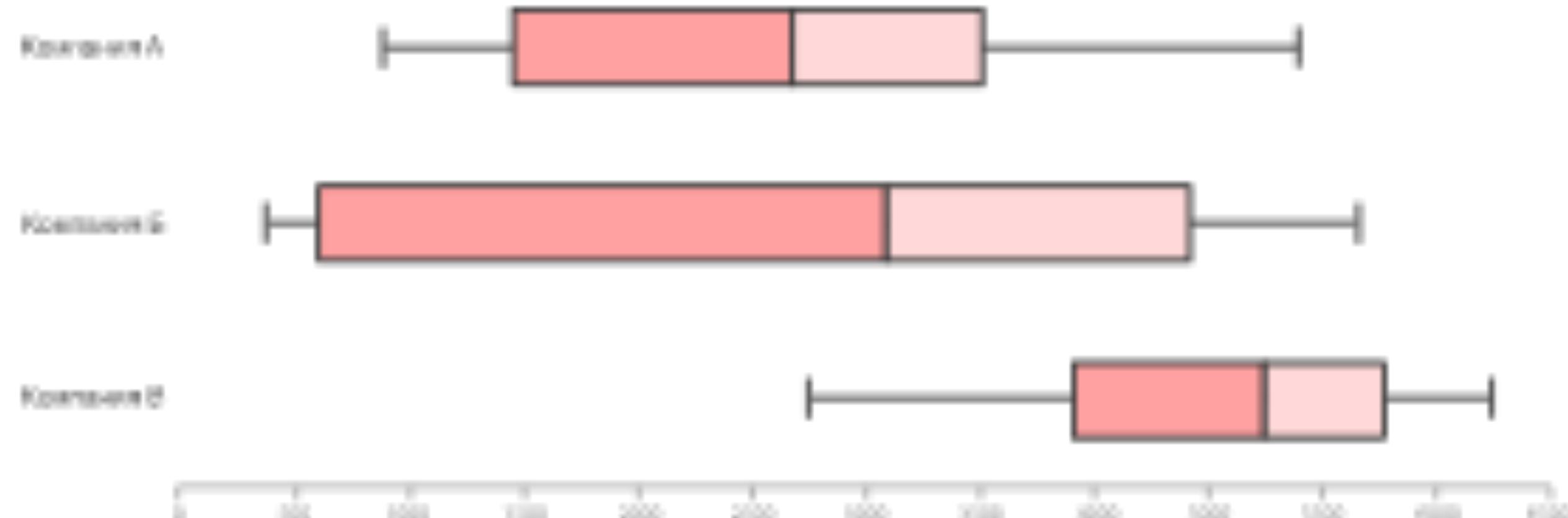
## Boxplot (Ящик с усами)

### Когда использовать?

- Для поиска выбросов и медианных значений в числовых данных.
- Для сравнения распределений между разными группами.

### Как интерпретировать?

- Медиана показывает центральное значение.
- Длинные усы говорят о большой изменчивости данных.
- Точки за пределами усов – это выбросы.



# Числовые данные

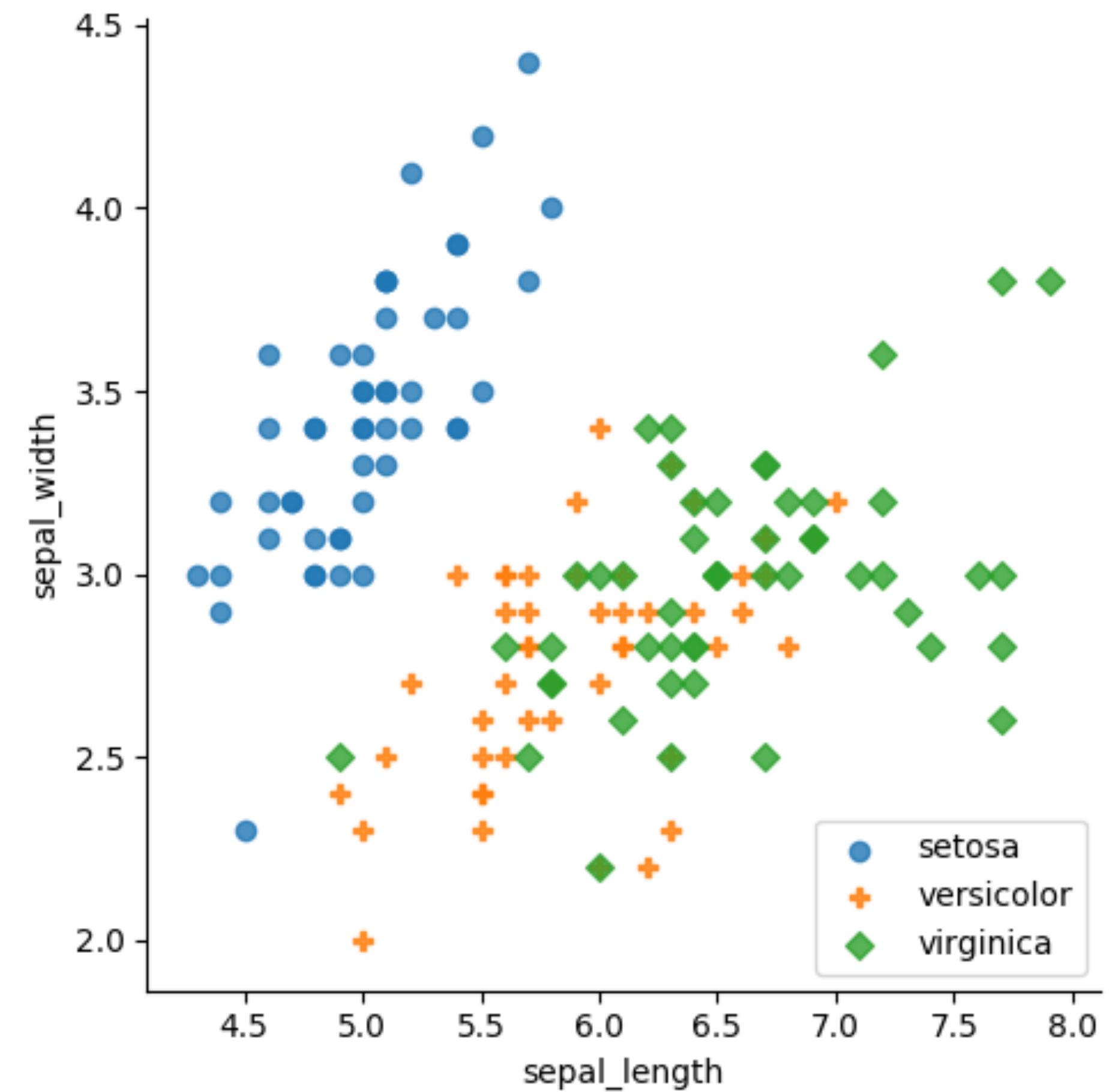
## Scatter plot (Точечная диаграмма)

### Когда использовать?

- Для анализа взаимосвязи между двумя числовыми переменными.

### Как интерпретировать?

- Если точки расположены вдоль диагонали, есть линейная зависимость.
- Если точки хаотично разбросаны, связи нет.
- Группирование точек может указывать на кластеры.



# Категориальные данные

**Категориальные данные** – это данные, разбитые на группы (например, регионы, продукты, профессии).

**Виды диаграмм:**

1. Bar chart
2. Heatmap
3. Tree map

# Категориальные данные

## Bar Chart (Столбчатая диаграмма)

### Когда использовать?

- Для сравнения категорий (например, продажи разных товаров).

### Как интерпретировать?

- Высокий столбец = большее значение.
- Упорядочивание столбцов по убыванию упрощает анализ.



# Категориальные данные

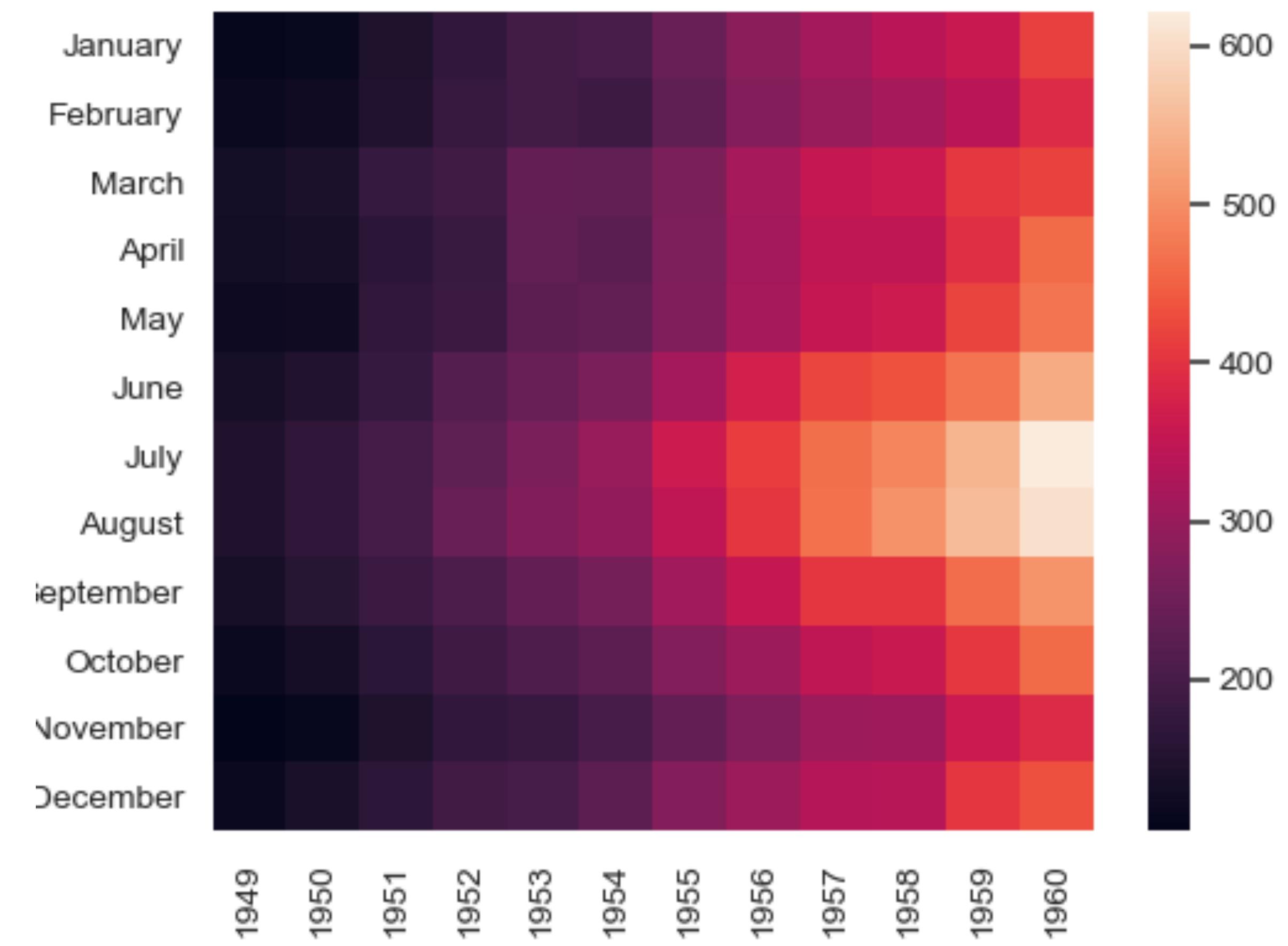
## Heatmap (Тепловая карта)

### Когда использовать?

- Для визуализации корреляции между переменными.
- Для анализа интенсивности значений в двухмерной таблице.

### Как интерпретировать?

- Чем насыщеннее цвет, тем выше значение.
- Светлые области – зоны с низкими значениями.



# Категориальные данные

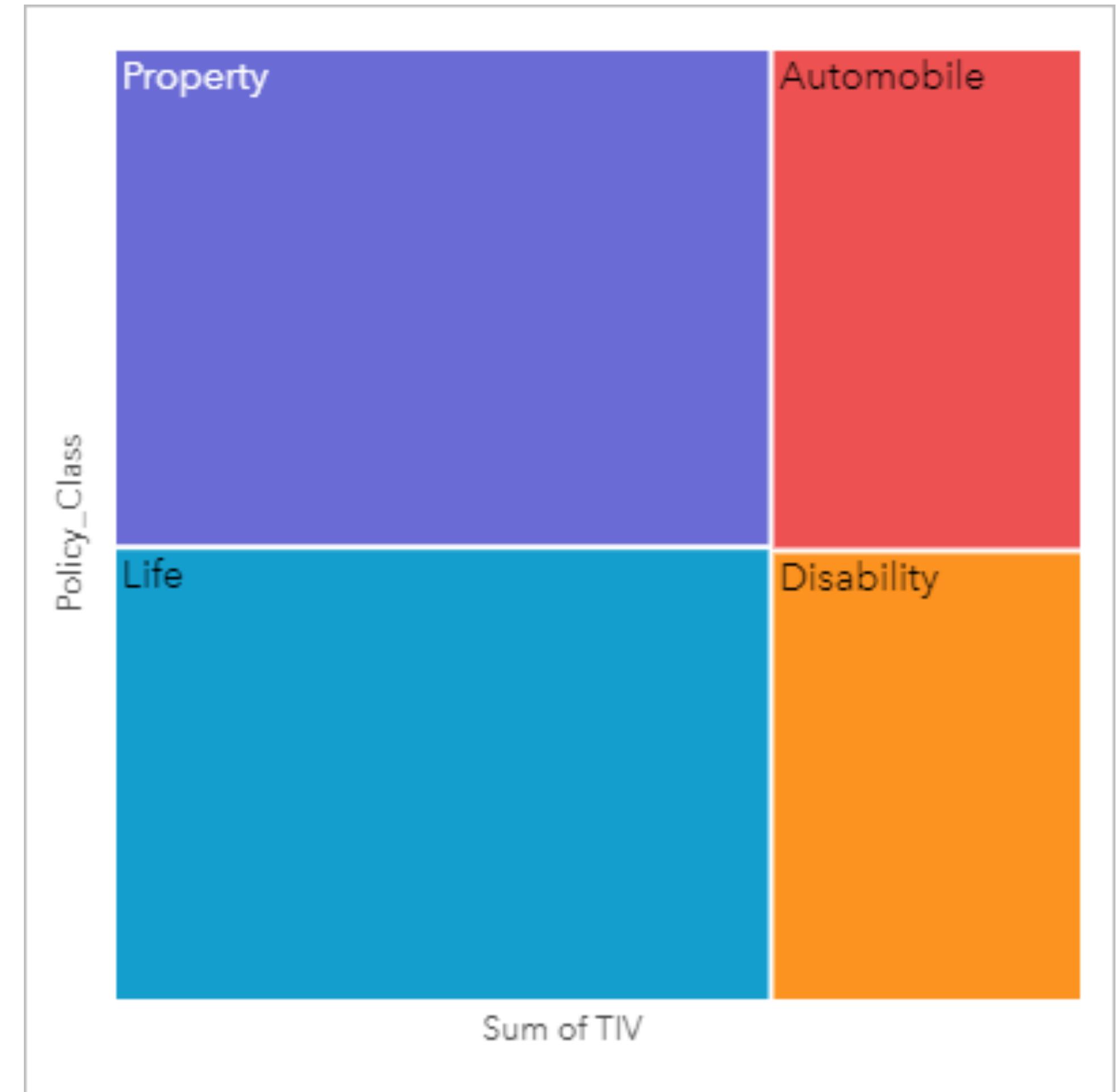
## Tree Map (Деревянная карта)

### Когда использовать?

- Когда важно показать вклад каждой категории в общее значение.

### Как интерпретировать?

- Чем больше блок, тем больше вклад категории.
- Цвет можно использовать для обозначения роста/спада.



# Временные ряды

**Временные ряды** – это данные, изменяющиеся во времени (например, продажи по дням).

**Виды диаграмм:**

1. Line chart
2. Area chart

# Временные ряды

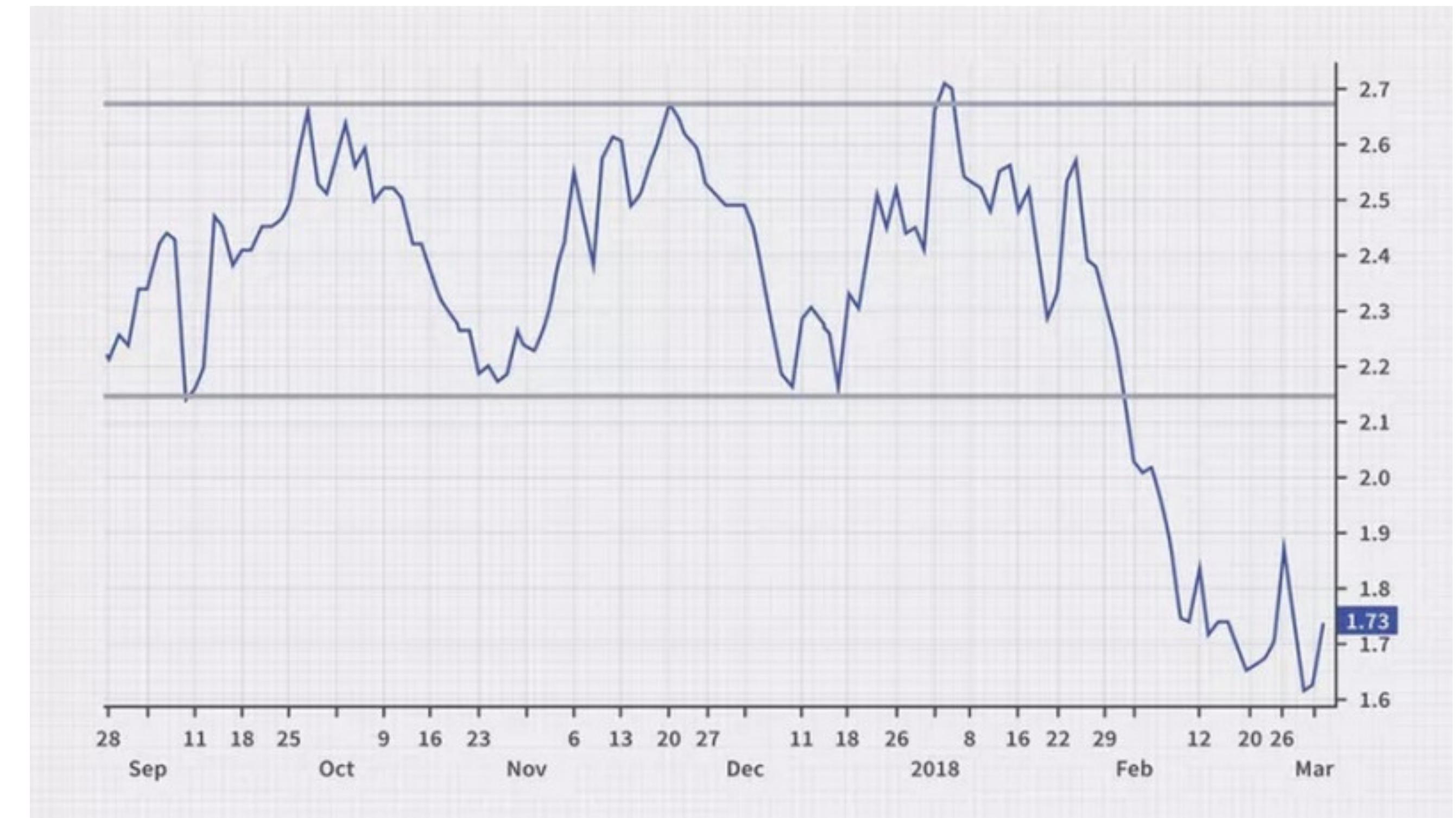
## Line Chart (Линейный график)

### Когда использовать?

- Для анализа трендов во времени.
- Для сравнения динамики нескольких показателей.

### Как интерпретировать?

- Восходящий тренд = рост.
- Спад = снижение.
- Резкие скачки могут указывать на аномалии.



# Временные ряды

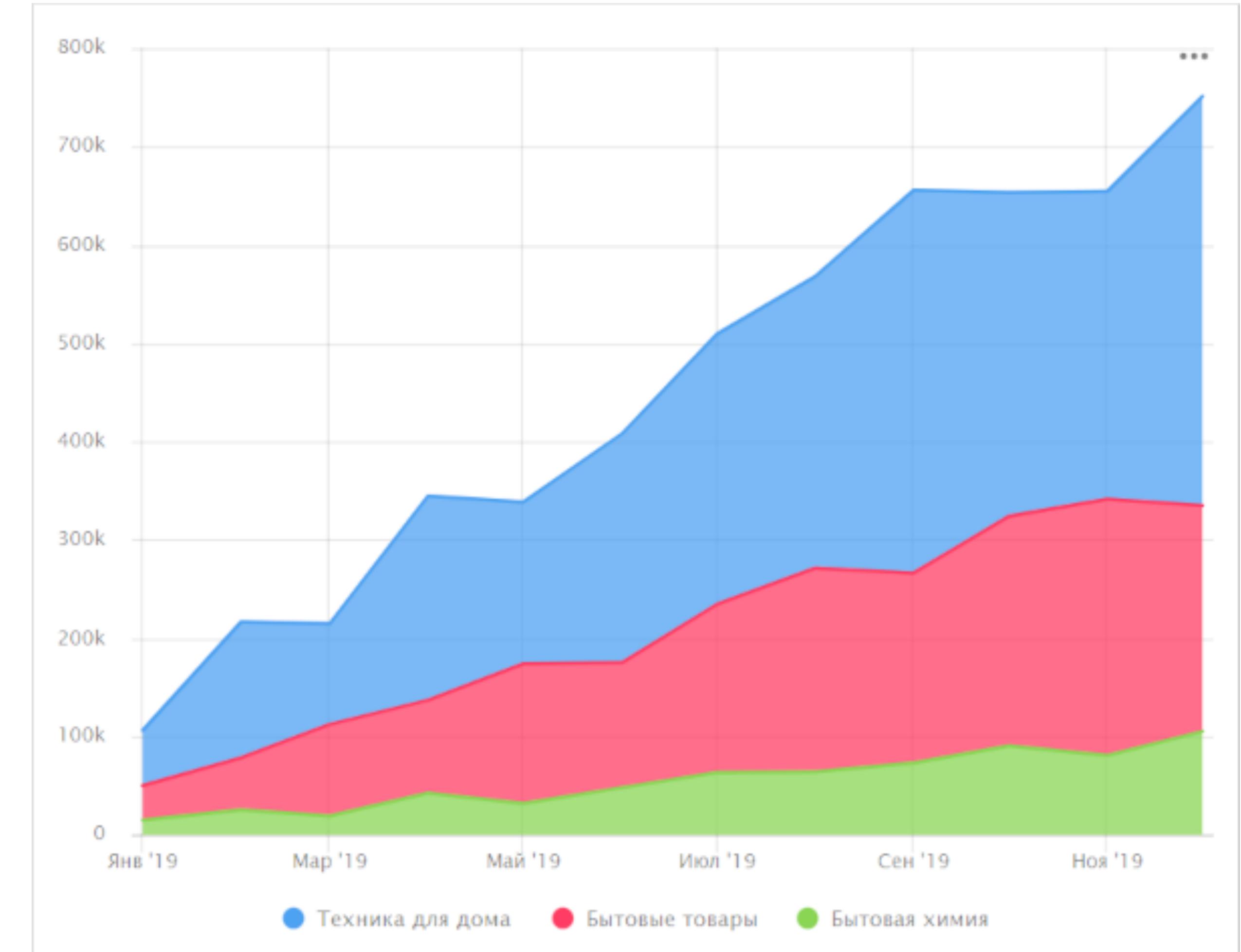
## Area Chart (Диаграмма областей)

### Когда использовать?

- Для демонстрации накопительных изменений во времени.

### Как интерпретировать?

- Чем больше заполненная область, тем выше значение.
- Хорошо показывает вклад разных категорий в общий тренд.



# **Геоданные: карты и их особенности**

**Геоданные** – это информация, привязанная к местоположению (например, распределение клиентов по странам).

**Виды карт:**

1. Карта плотности
2. Карта пузырьков

# Геоданные: карты и их особенности

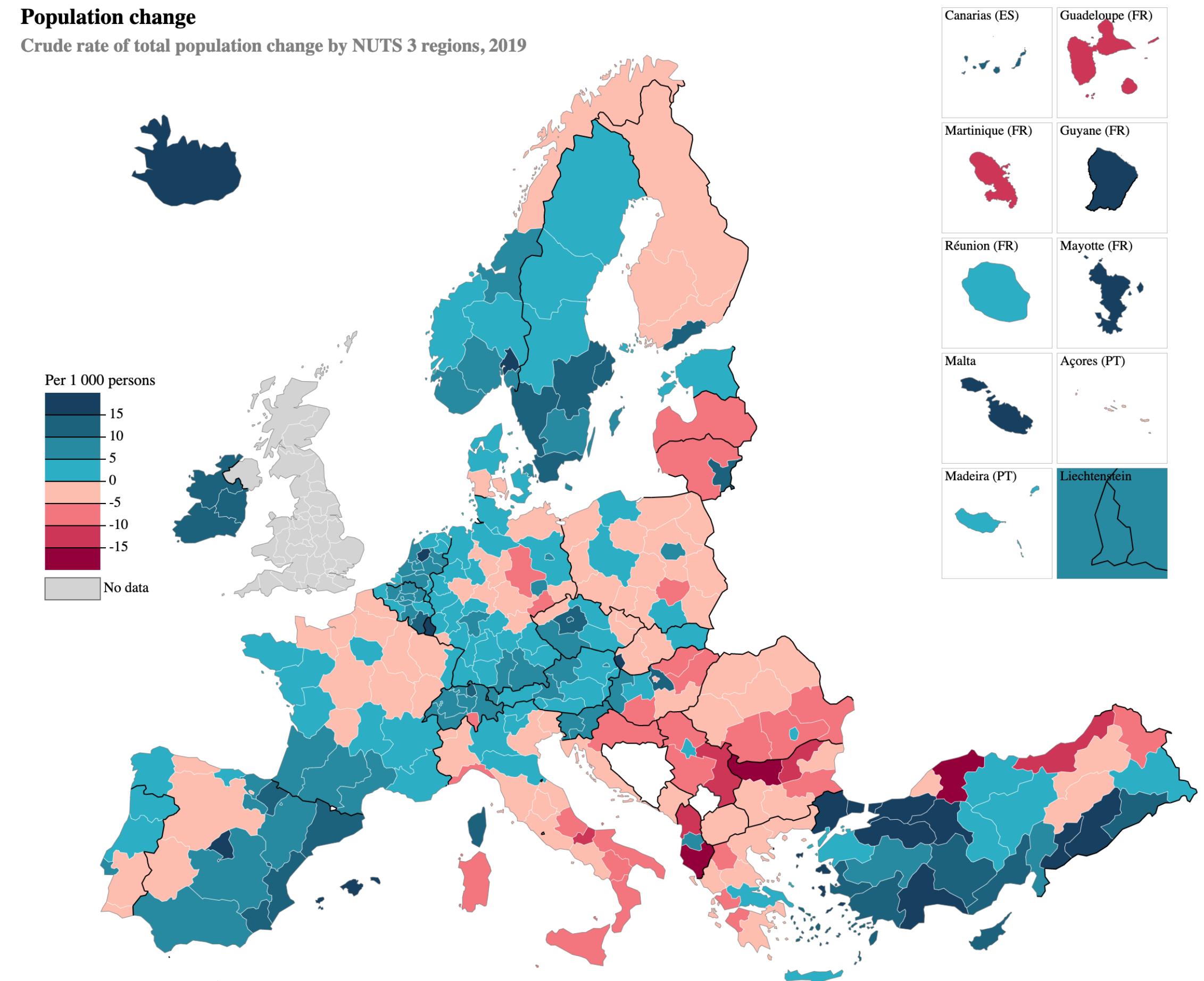
## Choropleth Map (Карта плотности)

### Когда использовать?

- Для сравнения числовых показателей по регионам.

### Как интерпретировать?

- Темнее цвет = выше значение.
- Светлые области = низкие показатели.



# Геоданные: карты и их особенности

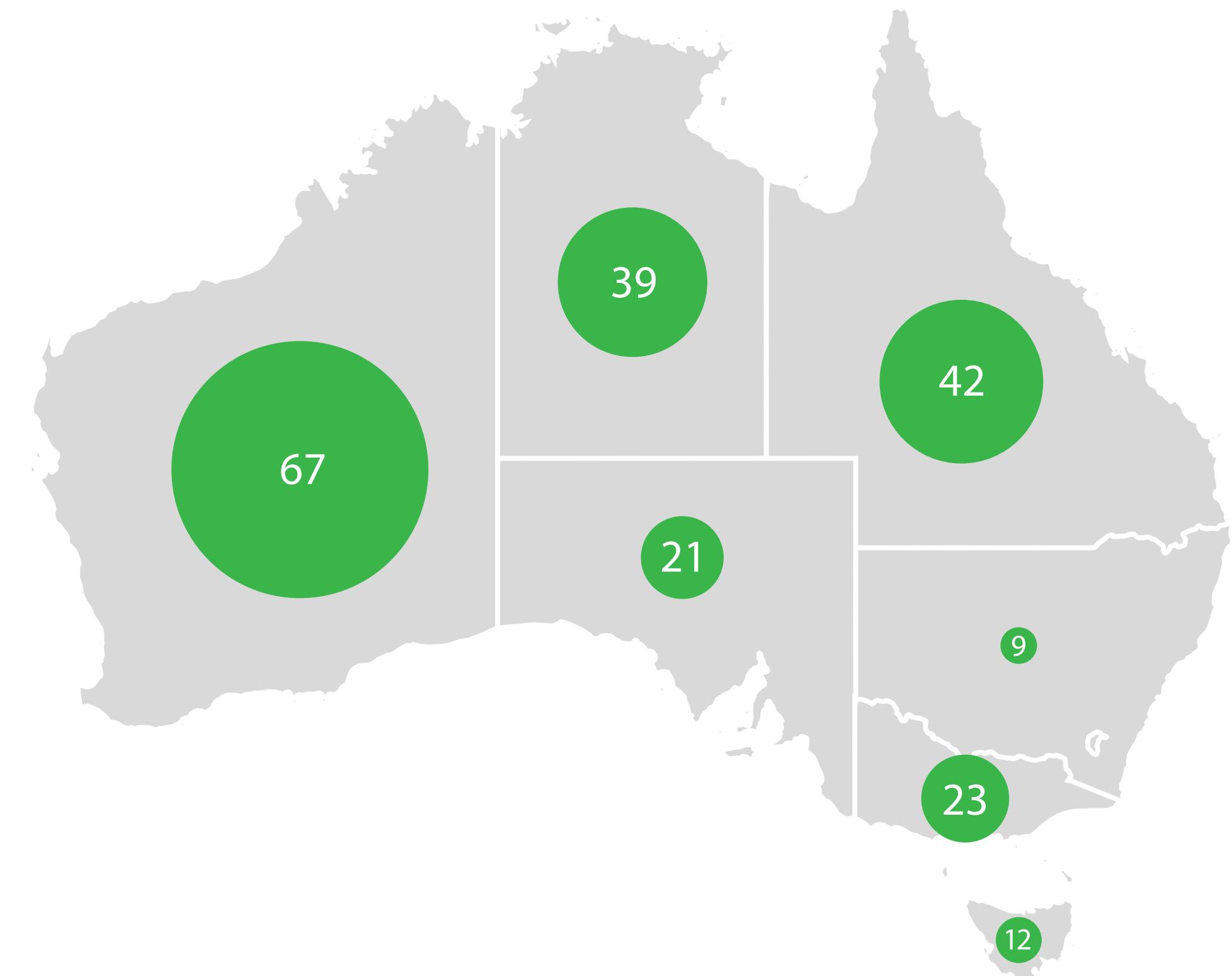
## Bubble Map (Карта пузырьков)

### Когда использовать?

- Когда важно показать размер значения в конкретной точке на карте.

### Как интерпретировать?

- Чем больше пузырек, тем выше значение.
- Цвет можно использовать для обозначения роста/спада.



# Вывод

1. **Числовые данные** → Гистограммы, Boxplot, Scatter Plot.
2. **Категориальные данные** → Bar Chart, Heatmap, Tree Map.
3. **Временные ряды** → Line Chart, Area Chart.
4. **Геоданные** → Choropleth Map, Bubble Map.

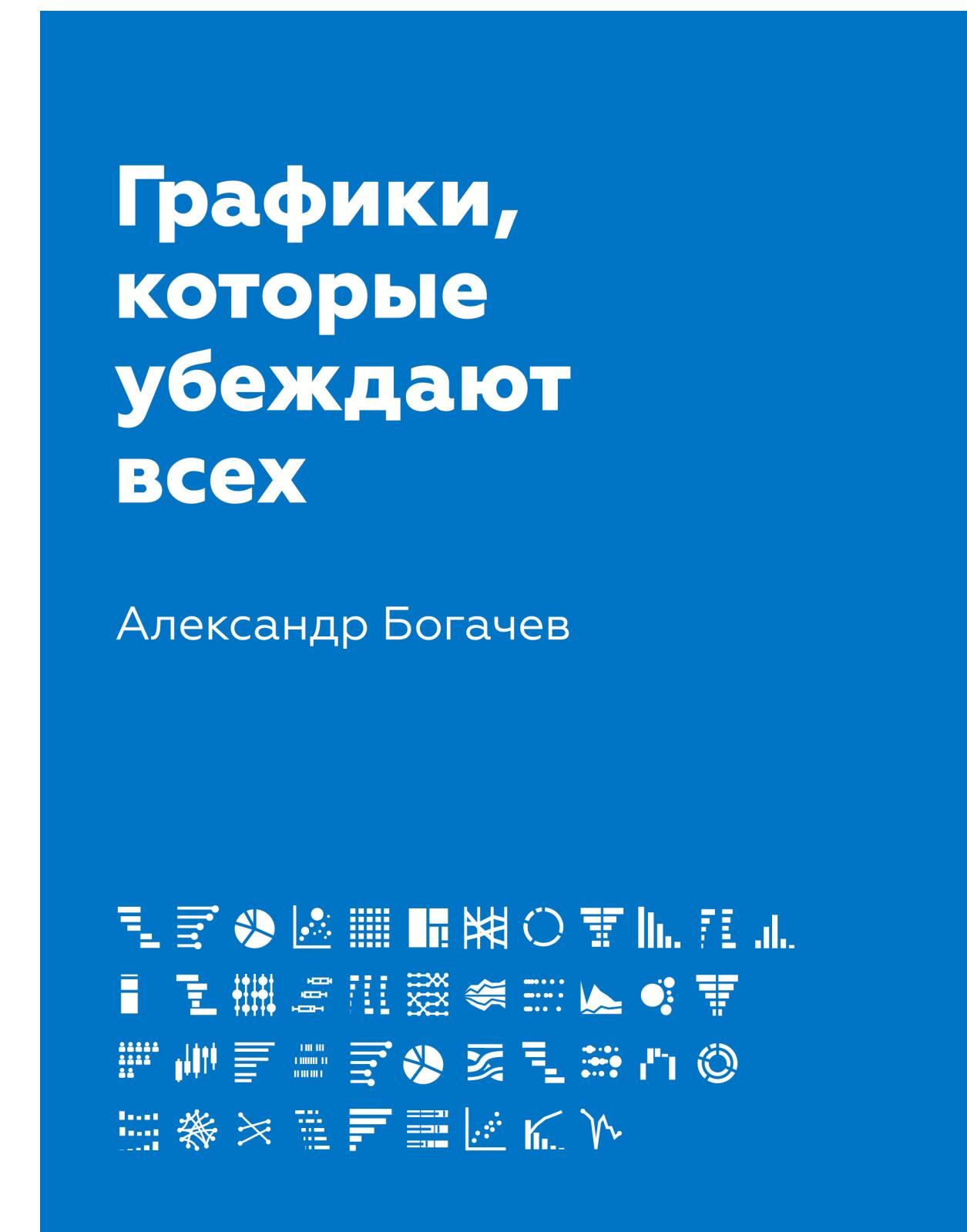
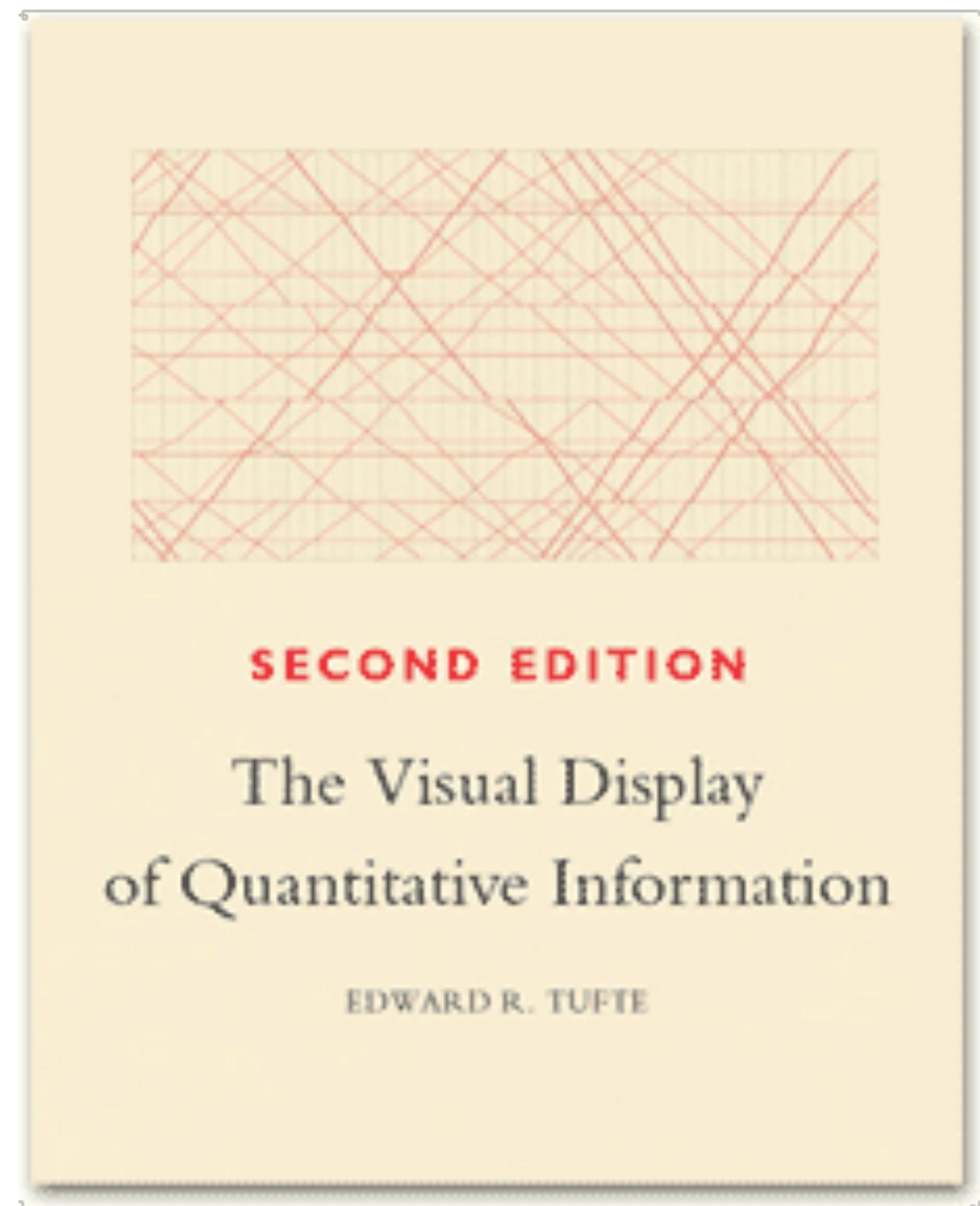
# Практика

Есть датасет `hr_data.csv`, в нем представлены данные о сотрудниках компании.

<https://clck.ru/3GgTBd>

Атрибут	Age (Возраст)	MonthlyIncome (Ежемесячный доход)	Department (Отдел)	Gender (Пол)	Education (Уровень образования)	Attrition (Текущесть кадров)
Описание	Возраст сотрудника в годах.	Уровень месячного дохода сотрудника в валюте компании.	Департамент, в котором работает сотрудник (Sales, Research & Development, HR и т. д.).	Пол сотрудника (Male – Мужчина, Female – Женщина)	Уровень образования сотрудника (1 – Среднее, 2 – Колледж, 3 – Бакалавр, 4 – Магистр, 5 – Доктор наук).	Показывает, покинул ли сотрудник компанию (Yes – уволился, No – остался в компании).

# Литература для доп погружения в тему



# Статьи для доп погружения в тему

1. О. В. Пескова «О визуализации информации». URL: <https://cyberleninka.ru/article/n/o-vizualizatsii-informatsii>
2. В. Пилюгин, Е. Маликова, А. Пасько, В. Аджиев «Научная визуализация как метод анализа научных данных». URL: <https://sv-journal.org/2012-4/06/index.html>
3. В. Л. Авербух, Д. В. Манаков «Анализ и визуализация больших данных». URL: <https://data.lact.ru/f1/s/o/299/basic/1605/962/033.pdf>
4. «30 лучших визуализаций данных 2024 года [с примерами]». URL: <https://visme.co/blog/ru/%Do% B2% Do% B8% Do% B7% D1% 83% Do% Bo% Do% BB% Do% B8% Do% B7% Do% Bo% D1% 86% Do% B8% D1% 8F-% Do% B4% Do% Bo% Do% BD% Do% BD% D1% 8B% D1% 85/>
5. «Визуализация данных: принципы, способы и полезные инструменты». URL: <https://goit.global/ua-ru/articles/vyzualyzatsiya-dann-kh-pryntsyp-sposob-y-polezn-e-ynstrument/>
6. «Инструменты визуализации данных: введение в BI-инструменты». URL: <https://digitalstrategy.ru/blog/vizualizaciya-dannyh-biznesa-osnovnye-instrumenty-i-principy/>

**Спасибо за внимание!**