20 МАЯ 19:00 МСК
ОНЛАЙН

РАЗРУШИТЕЛИ СТАТИСТИЧЕСКИХ МИФОВ
2 СЕЗОН

ЕЛЕНА УБОГОЕВА | МИФ №6:
ВИЗУАЛИЗАЦИЯ – ЭТО ПРОСТО КРАСИВЫЕ ГРАФИКИ

HTTPS://T.ME/CHAT_BIOSTAT_R

План лекции

- Зачем нужна визуализация?
- Как не надо делать графики
  - Типичные ошибки в визуализации данных
  - Интерпретация боксплотов
  - Интерпретация пределов погрешностей на графиках (стандартное отклонение, стандартная ошибка, доверительный интервал)
- Как надо делать графики
  - Правила Эдварда Тафти

Разные цели графиков

Графики можно делать для разных задач:

- Для себя, посмотреть на данные (exploratory data analysis, EDA)
- Для представления результатов в публикации, презентации, отчете

В данной лекции речь пойдет в основном о втором виде

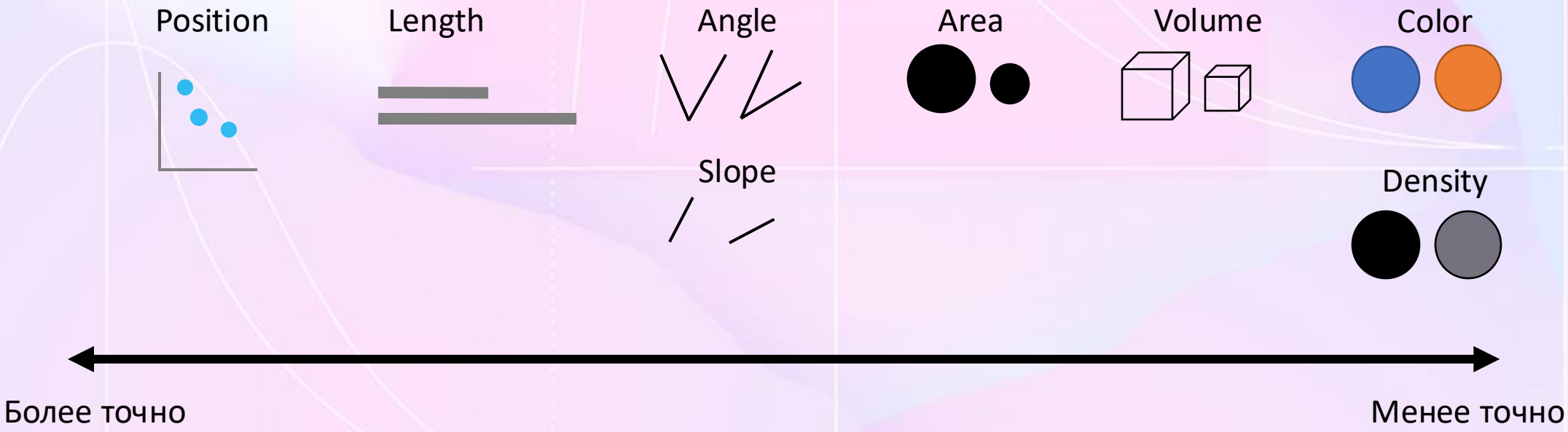
Зачем нужна визуализация?

Два подхода:

- Данные объективны сами по себе. Задача ученого/аналитика — "услышать" данные, минимизируя субъективное влияние.

- Данные сами по себе ничего не говорят. Важно как мы анализируем, представляем данные, какую модель используем.

Кодирование данных в визуализации



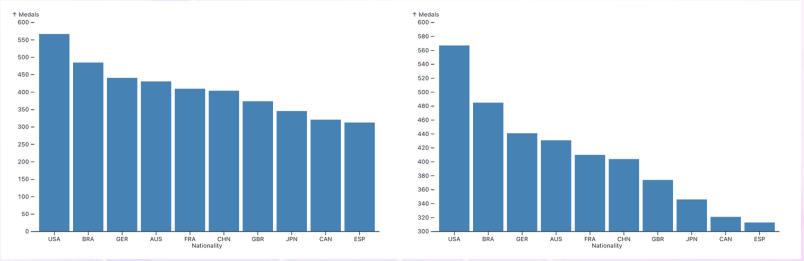
Как не надо делать графики

Типичные ошибки в визуализации данных

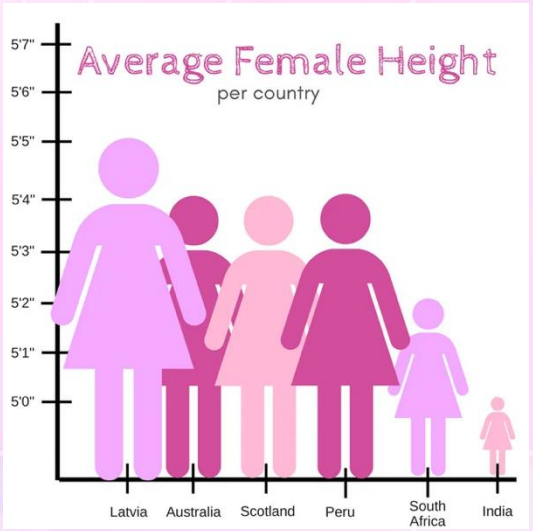
Ошибка 1	Ошибка 2	Ошибка 3
Ошибка 4	Ошибка 5	Ошибка 6

Типичные ошибки в визуализации данных

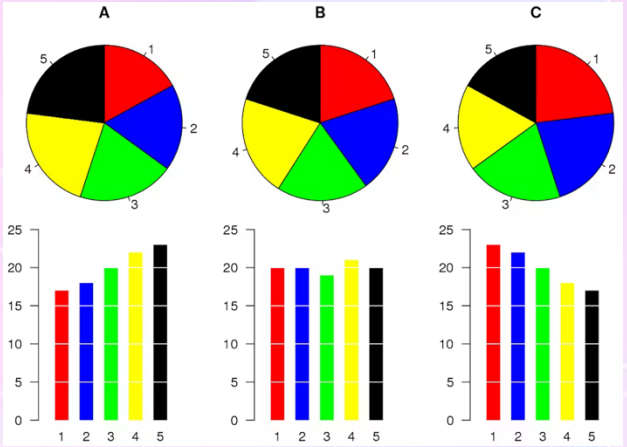
Axis cropping



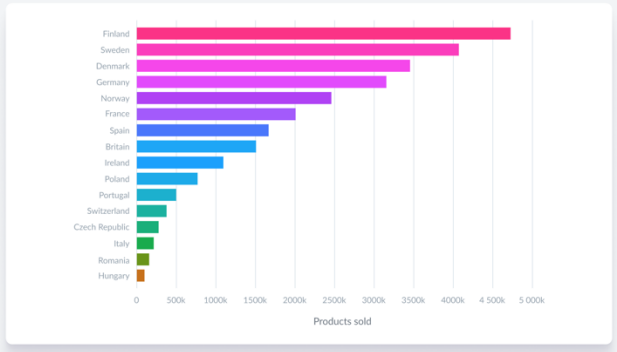
Использование объема



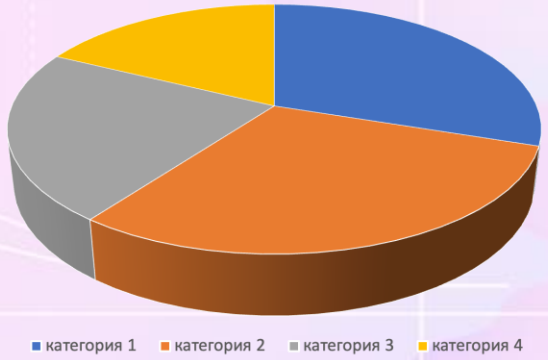
Использование pie chart для сравнения категорий



Бессмысленное использование цветов

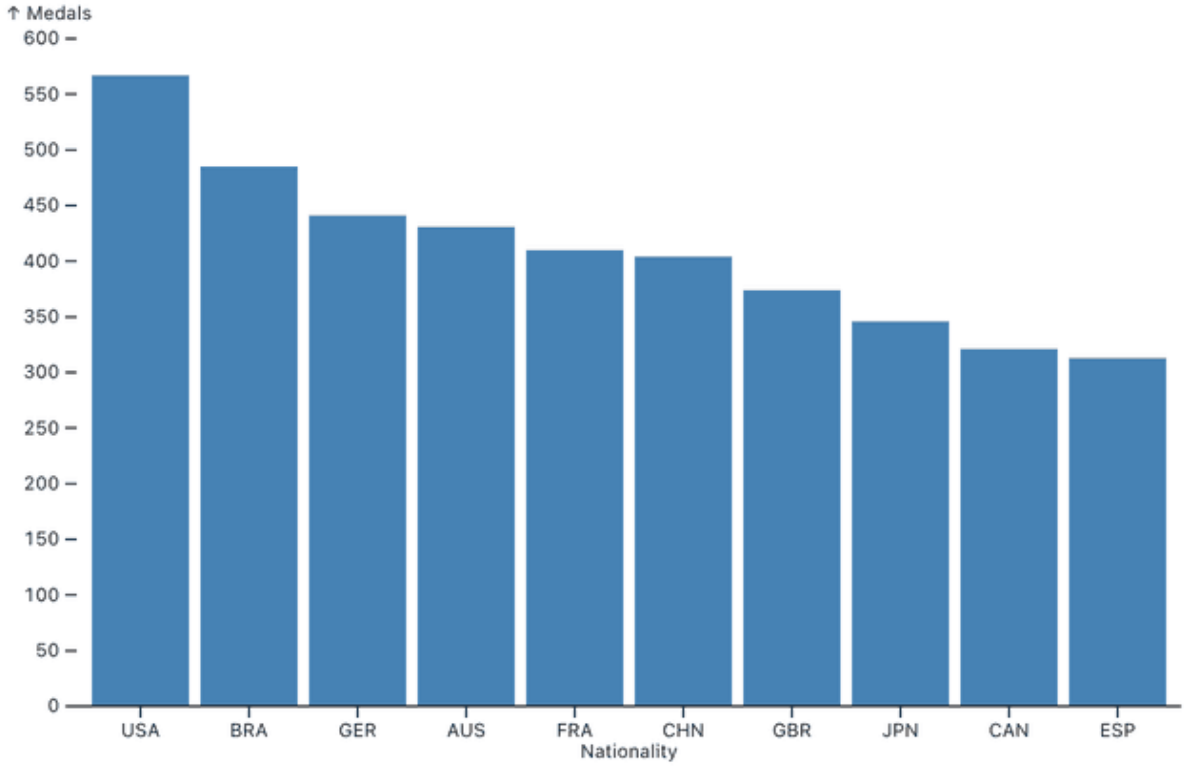


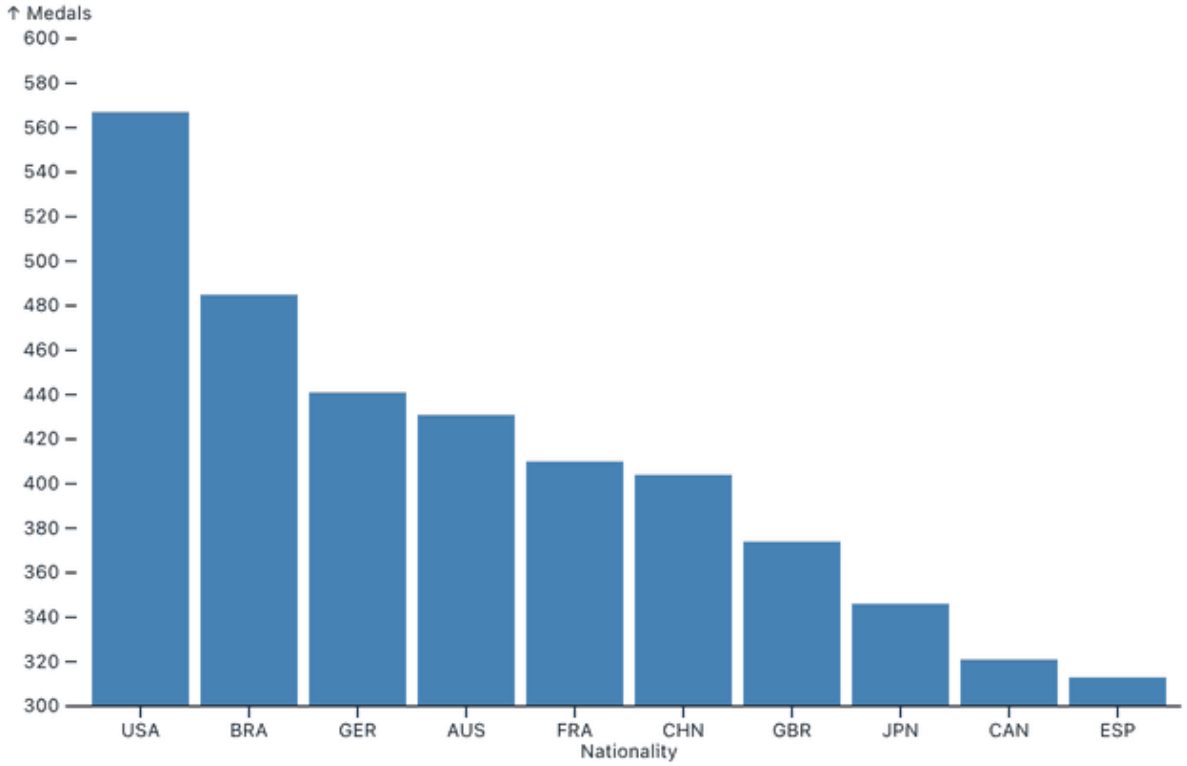
Неправильное использование 3D



Неправильный выбор типа графика

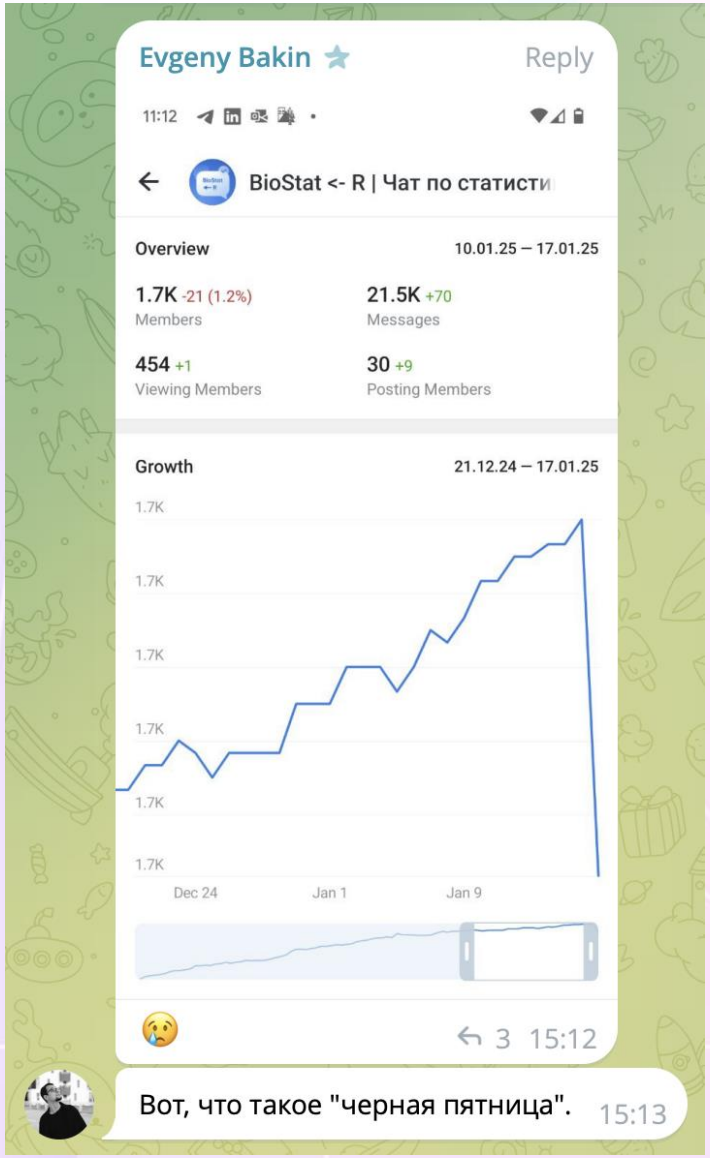
Axis cropping





Ось начинается не с нуля, в результате различия кажутся больше, чем они есть на самом деле

Axis cropping



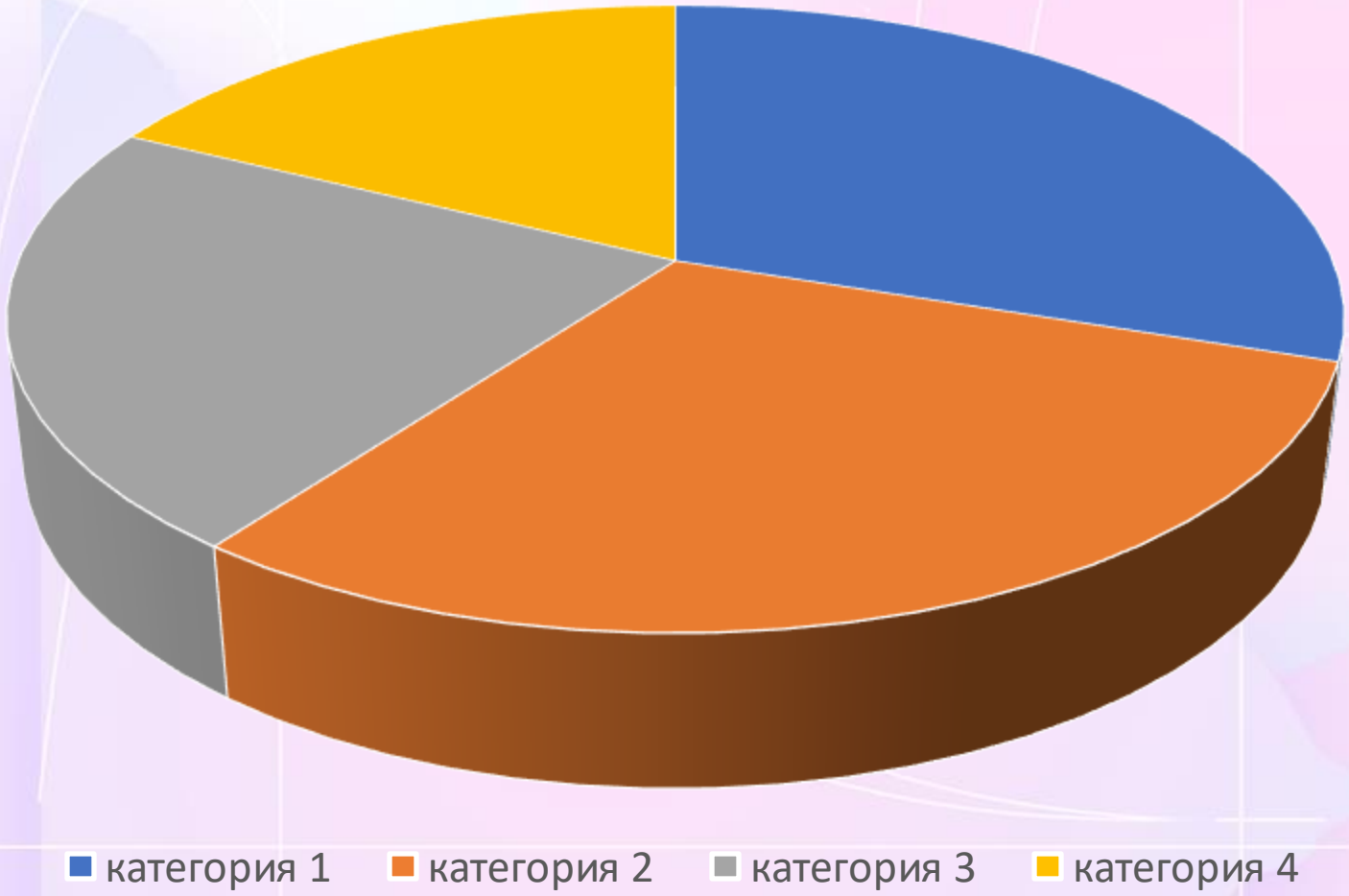
Однако в редких случаях отрывать ось от нуля можно.

Это допустимо, в случае, если нас интересует тренд, а не абсолютные значения.

Axis cropping: пример когда это не ошибка

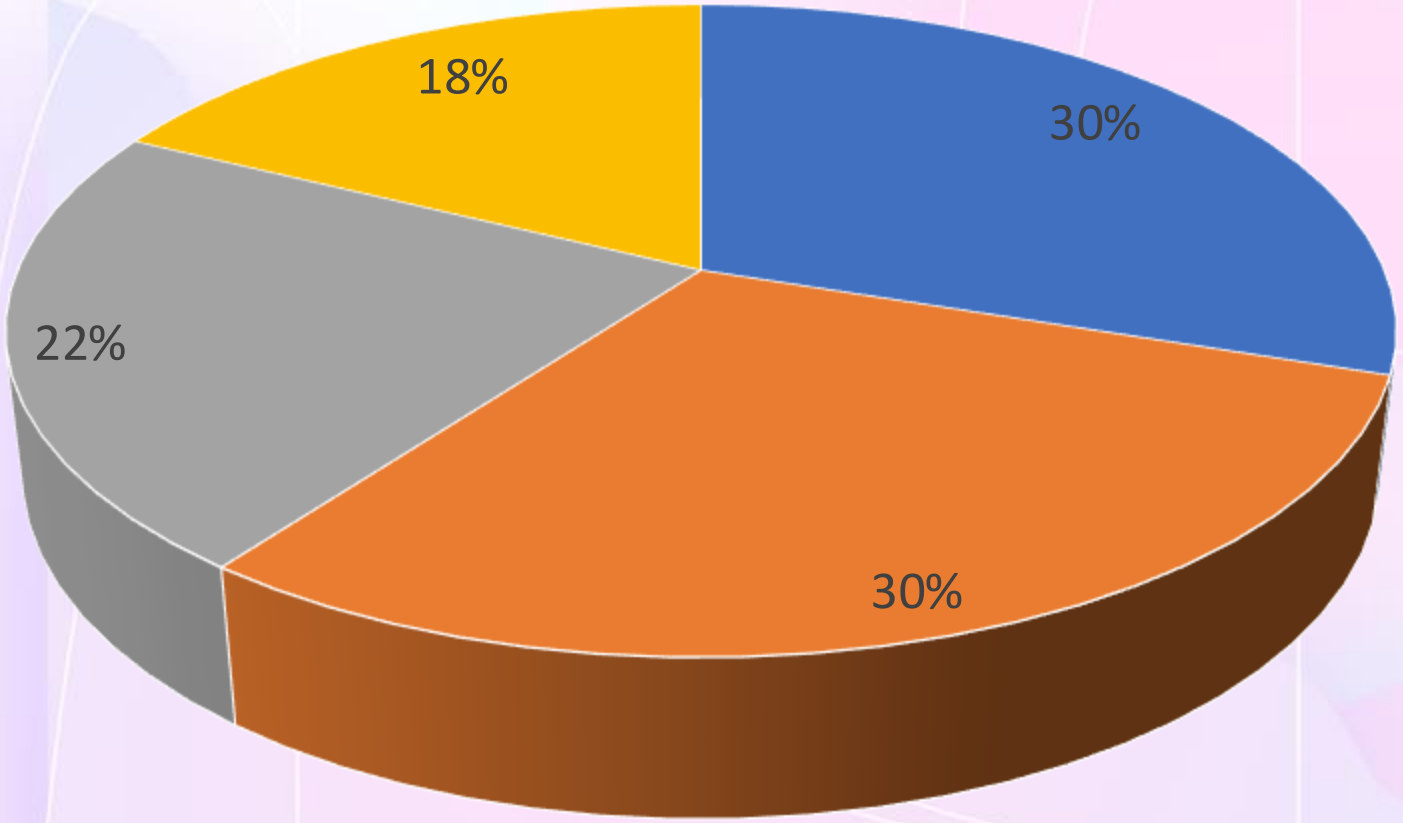

```
library(patchwork)
trend <- 0.95 + (0.02 * (1:100 - 1) / (100 - 1))
set.seed(42)
noise <- rnorm(100, mean = 0, sd = 0.005)
quality <- trend + noise

df <- data.frame(dt = dt, quality = quality)

plot1 <- df %>%
  ggplot(aes(dt, quality, group = 1))+
  geom_line()+
  ggtitle('Если отрывать ось от нуля')+
  theme_minimal()

plot2 <- df %>%
  ggplot(aes(dt, quality, group = 1))+
  geom_line()+
  ggtitle('Если не отрывать ось от нуля')+
  scale_y_continuous(limits = c(0,1))+
  theme_minimal()

plot1 / plot2
```

3D пайчарт – хрестоматийный пример плохой визуализации



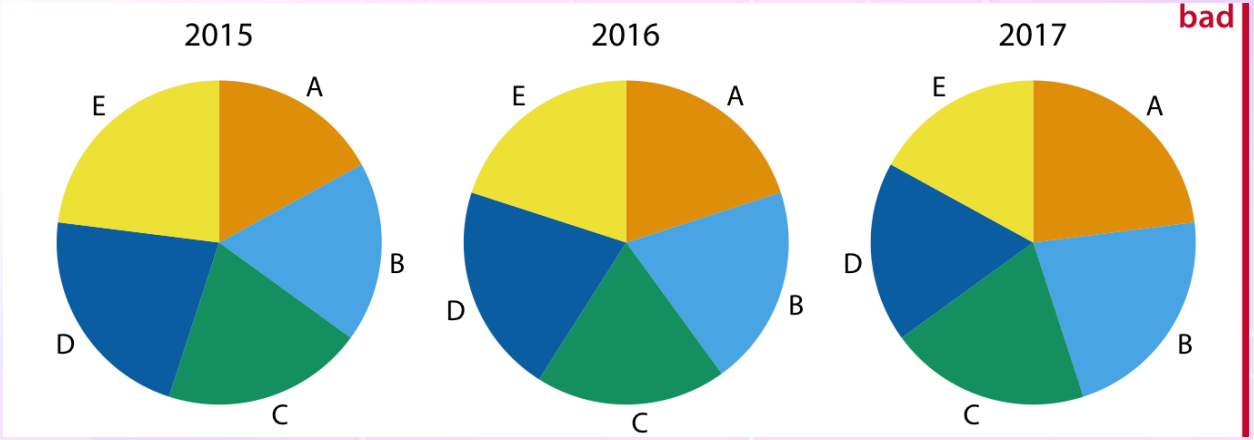
3D искажает пропорции из-за эффекта перспективы, в результате чего доли вблизи кажутся больше

3D пайчарт – хрестоматийный пример плохой визуализации



3D искажает пропорции из-за эффекта перспективы, в результате чего доли вблизи кажутся больше

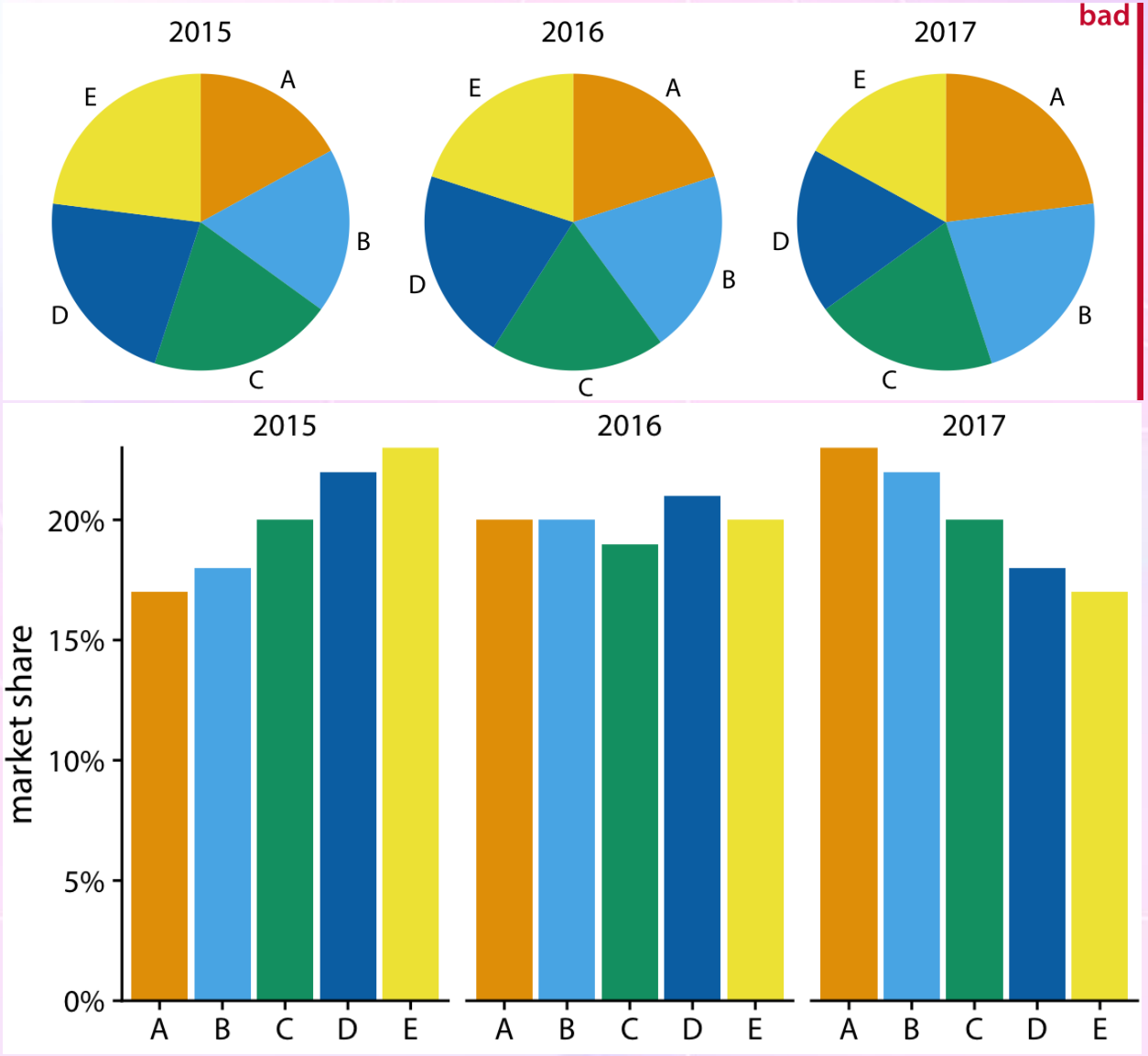
Пайчарт для сравнения категорий: тоже плохо



Как изменилась динамика компаний B и C?

https://clauswilke.com/dataviz/visualizing-proportions.html

Пайчарт для сравнения категорий: тоже плохо

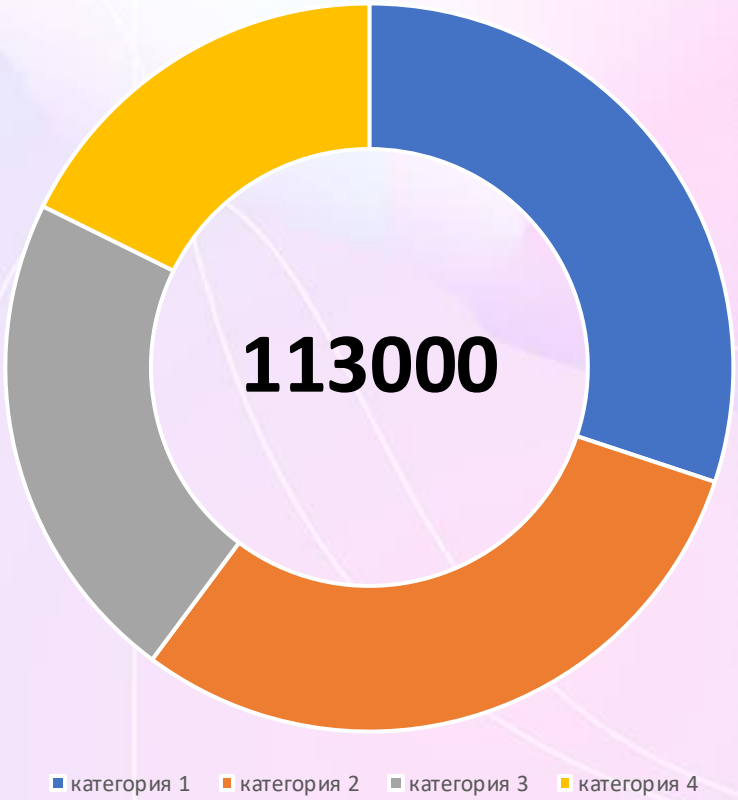


Сравнить динамику групп на барплоте в разы легче, чем с помощью pie chart

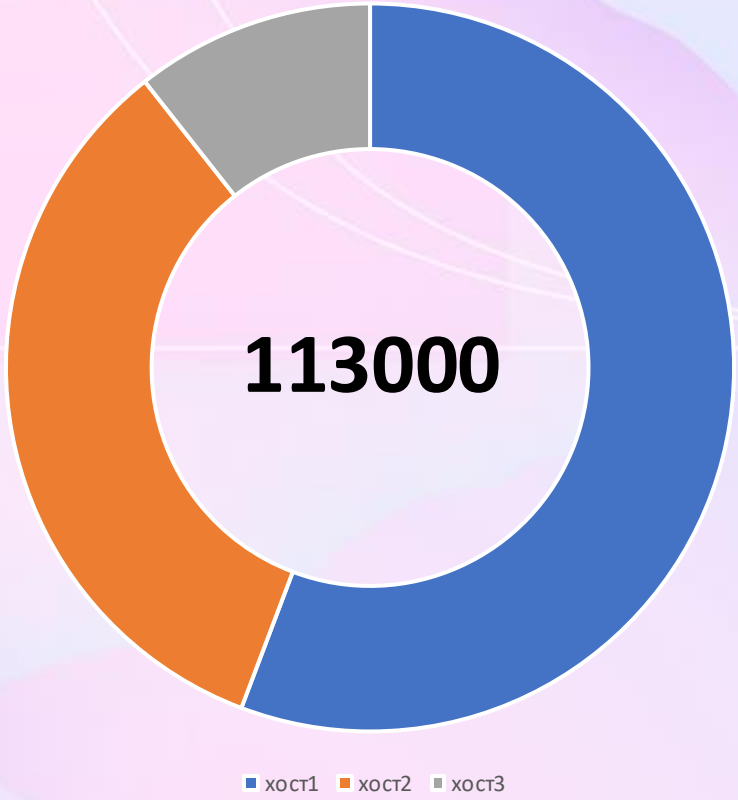
https://clauswilke.com/dataviz/visualizing-proportions.html

Пайчарт не всегда плохо

Сумма продаж по категориям



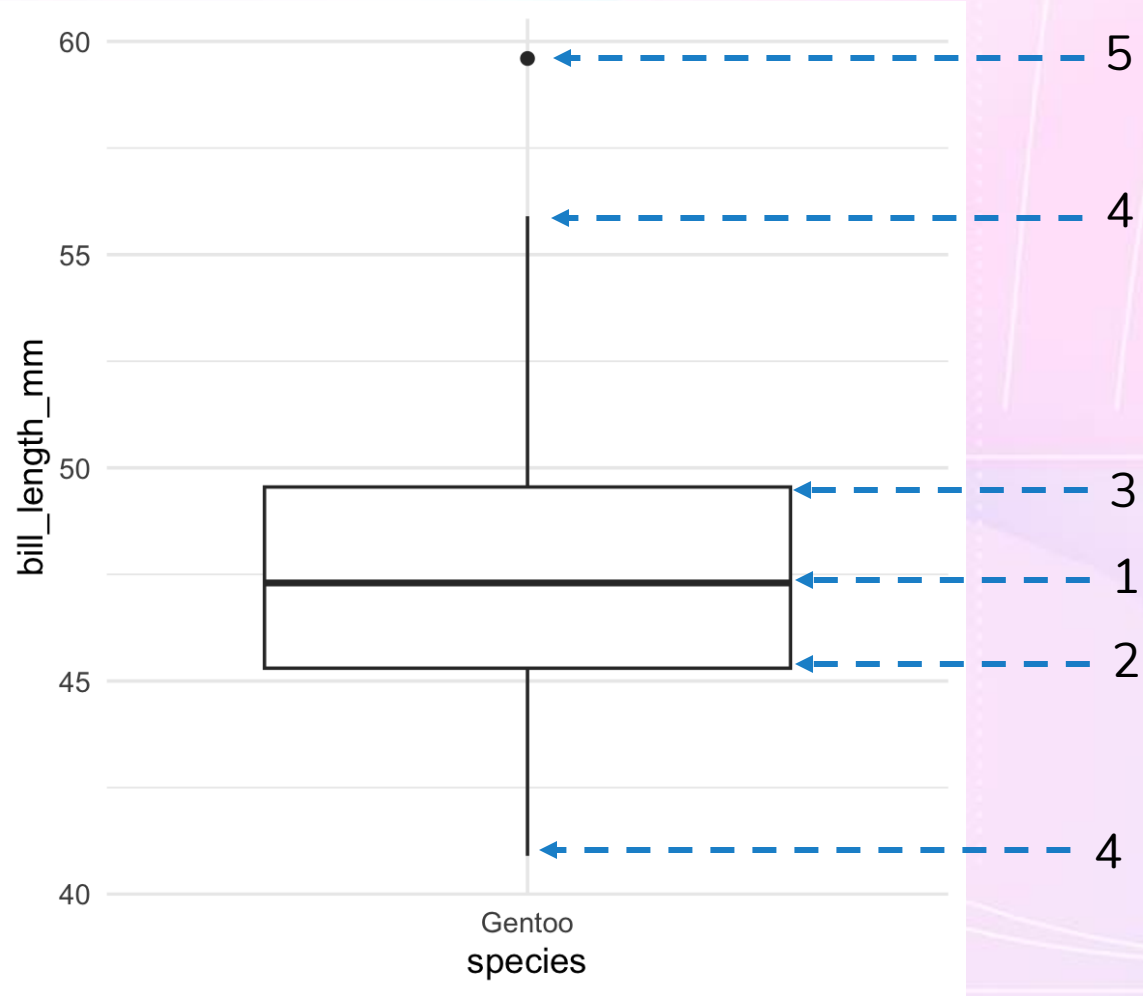
Сумма продаж по хостам



При этом если нет задачи сравнивать категории, то pie chart/donut chart нормально справляются

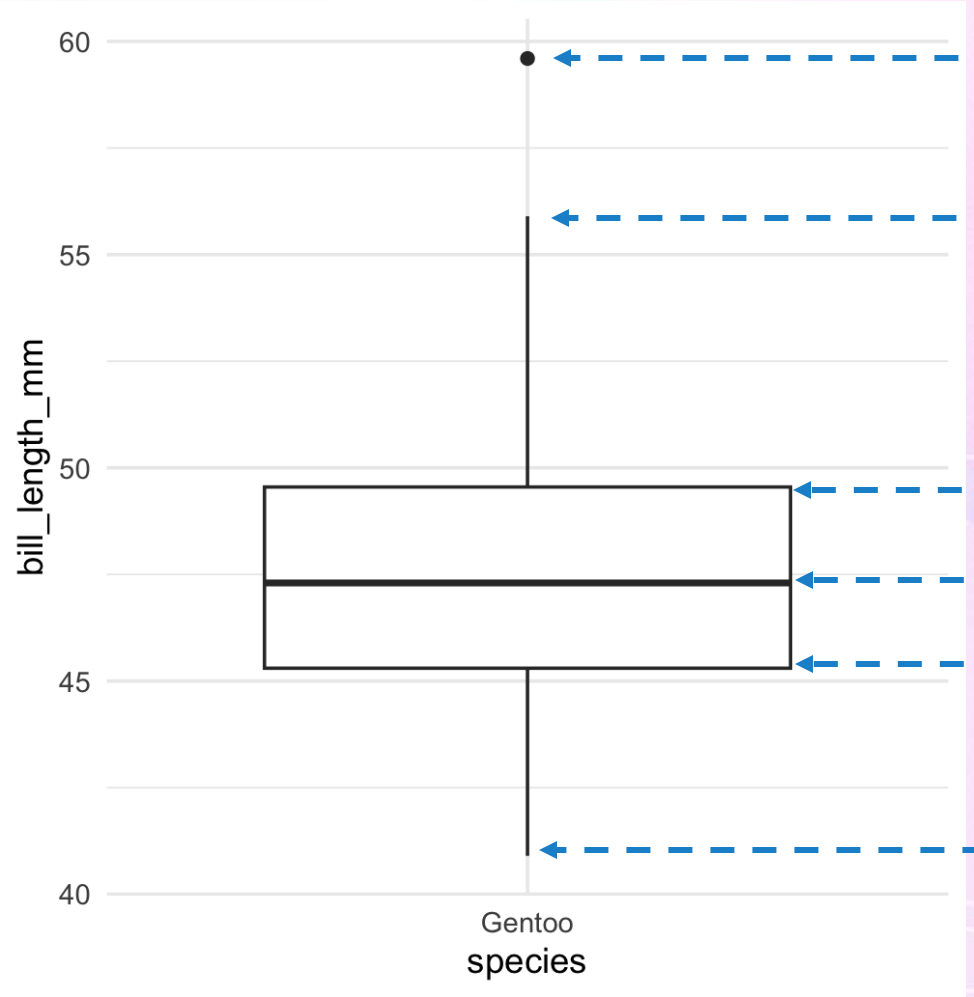
Интерпретация боксплотов

~~Что в ящике?~~
Что означает боксплот?




```
library(tidyverse)
library(palmerpenguins)

penguins %>%
  filter(species == 'Gentoo') %>%
  ggplot(aes(species,
bill_length_mm))+
  geom_boxplot()+
  theme_minimal()
```

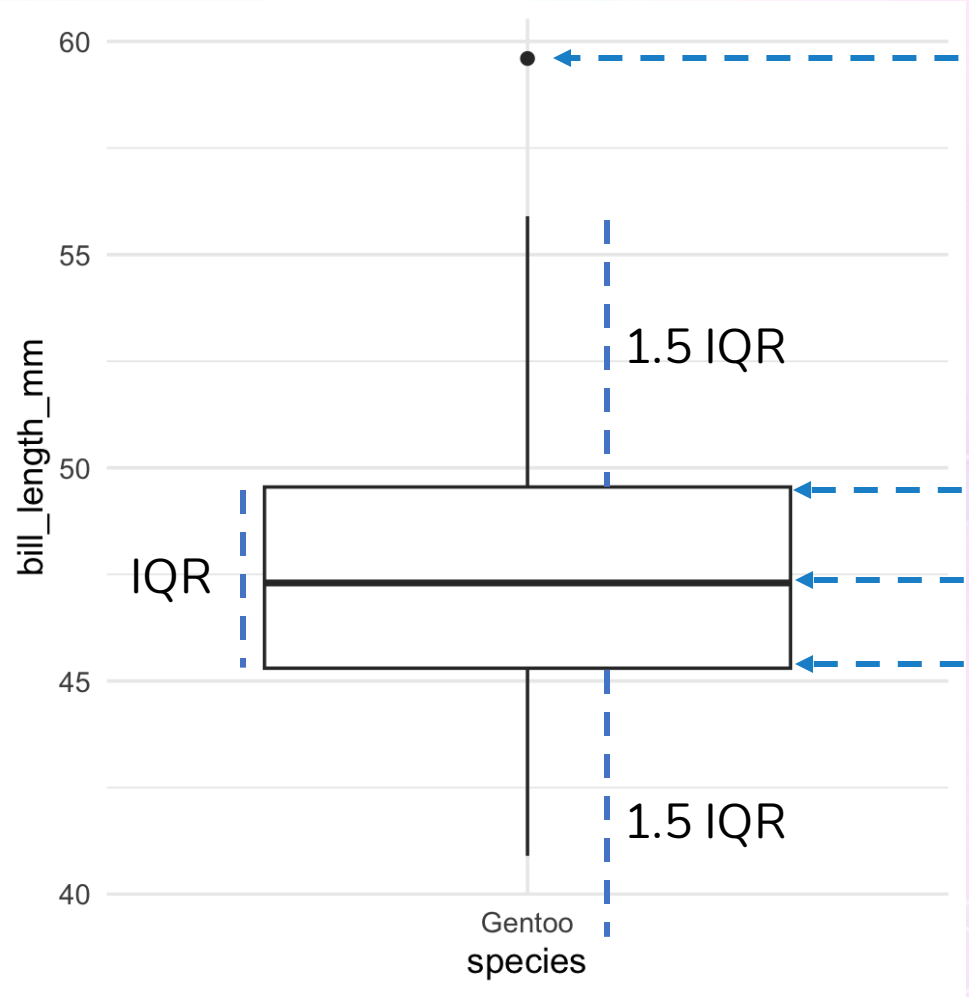
~~Что в ящике?~~
Что означает боксплот?




```
library(tidyverse)
library(palmerpenguins)

penguins %>%
  filter(species == 'Gentoo') %>%
  ggplot(aes(species,
bill_length_mm))+
  geom_boxplot()+
  theme_minimal()
```

Усы боксплота соответствуют наблюдаемым максимальным и минимальным значениям от соответствующего квартиля в пределах 1.5 IQR

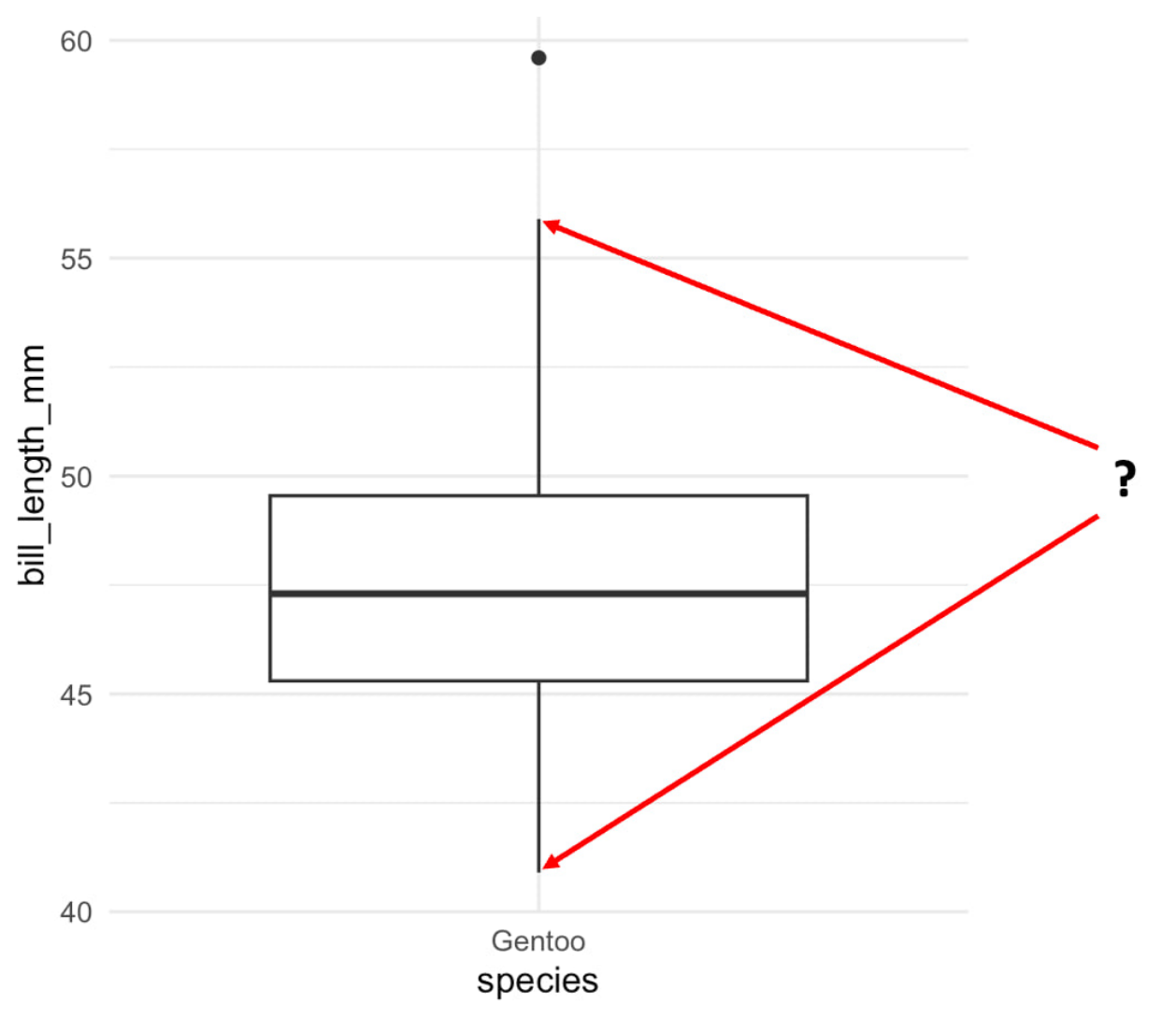
~~Что в ящике?~~
Что означает боксплот?




```
library(tidyverse)
library(palmerpenguins)

penguins %>%
  filter(species == 'Gentoo') %>%
  ggplot(aes(species,
bill_length_mm))+
  geom_boxplot()+
  theme_minimal()
```

Усы боксплота соответствуют наблюдаемым максимальным и минимальным значениям от соответствующего квартиля в пределах 1.5 IQR

Что думают подписчики о боксплоте



Статистика и R в науке и аналитике

Что отражают усы (whiskers) на графике box-plot?

Anonymous Poll

5% Стандартное отклонение (sd)

1% Стандартную ошибку (se)

7% Доверительный интервал (CI)

12% Минимум и максимум

4% 10% и 90% данных

19% Верхний и нижний квартили (Q1, Q3)

8% Межквартильный размах (IQR)

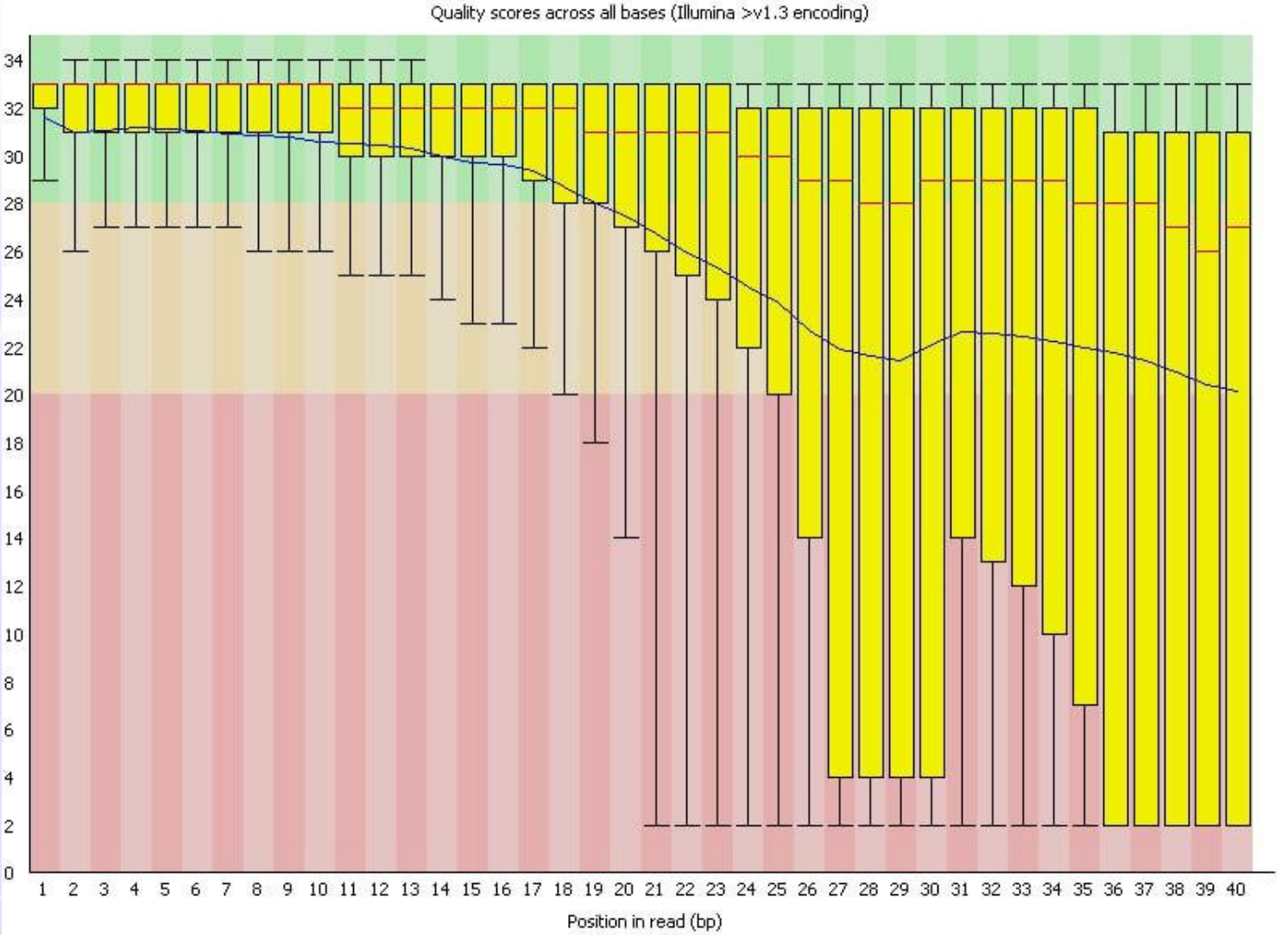
31% Наблюдаемые максимальные и минимальные значения от соответствующего квартиля в пределах 1.5 IQR

9% Нет однозначного ответа

4% Не знаю

864 votes

Боксплот может быть разным



The upper and lower whiskers represent the 10% and 90% points

https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3%20Analysis%20Modules/2%20Per%20Base%20Sequence%20Quality.html

Боксплот может быть разным

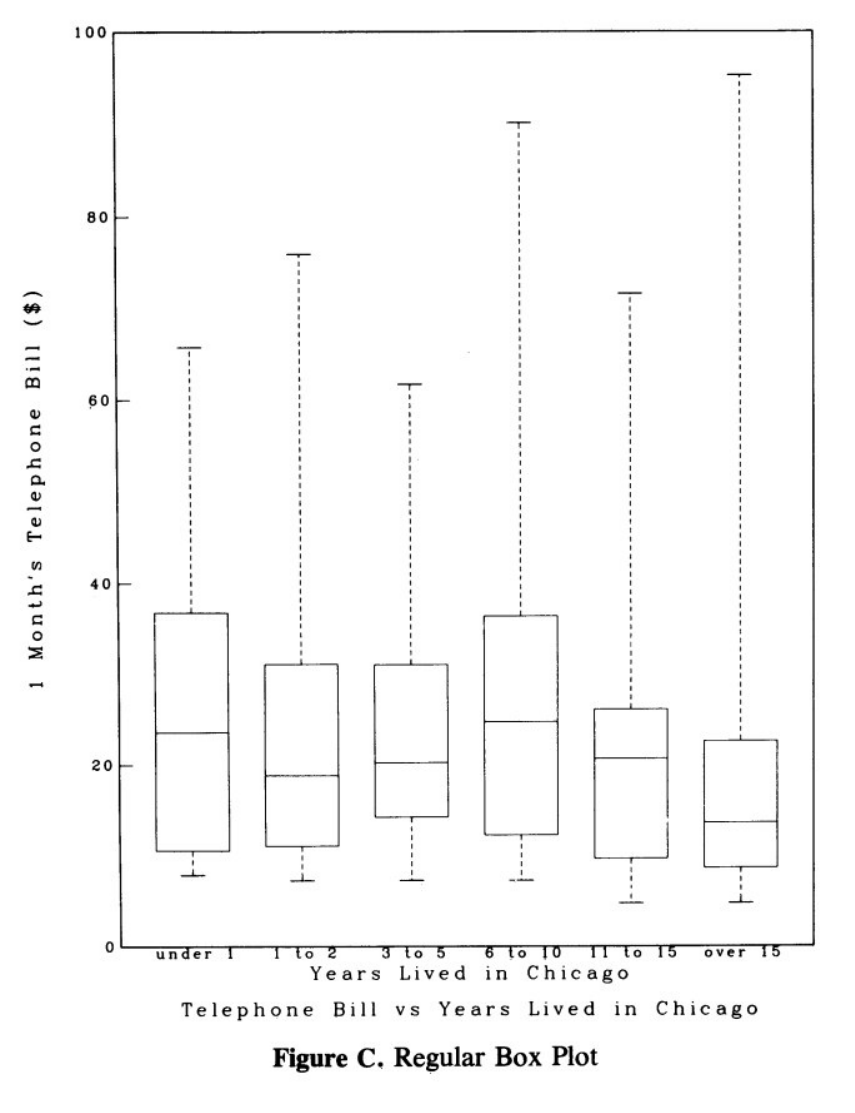


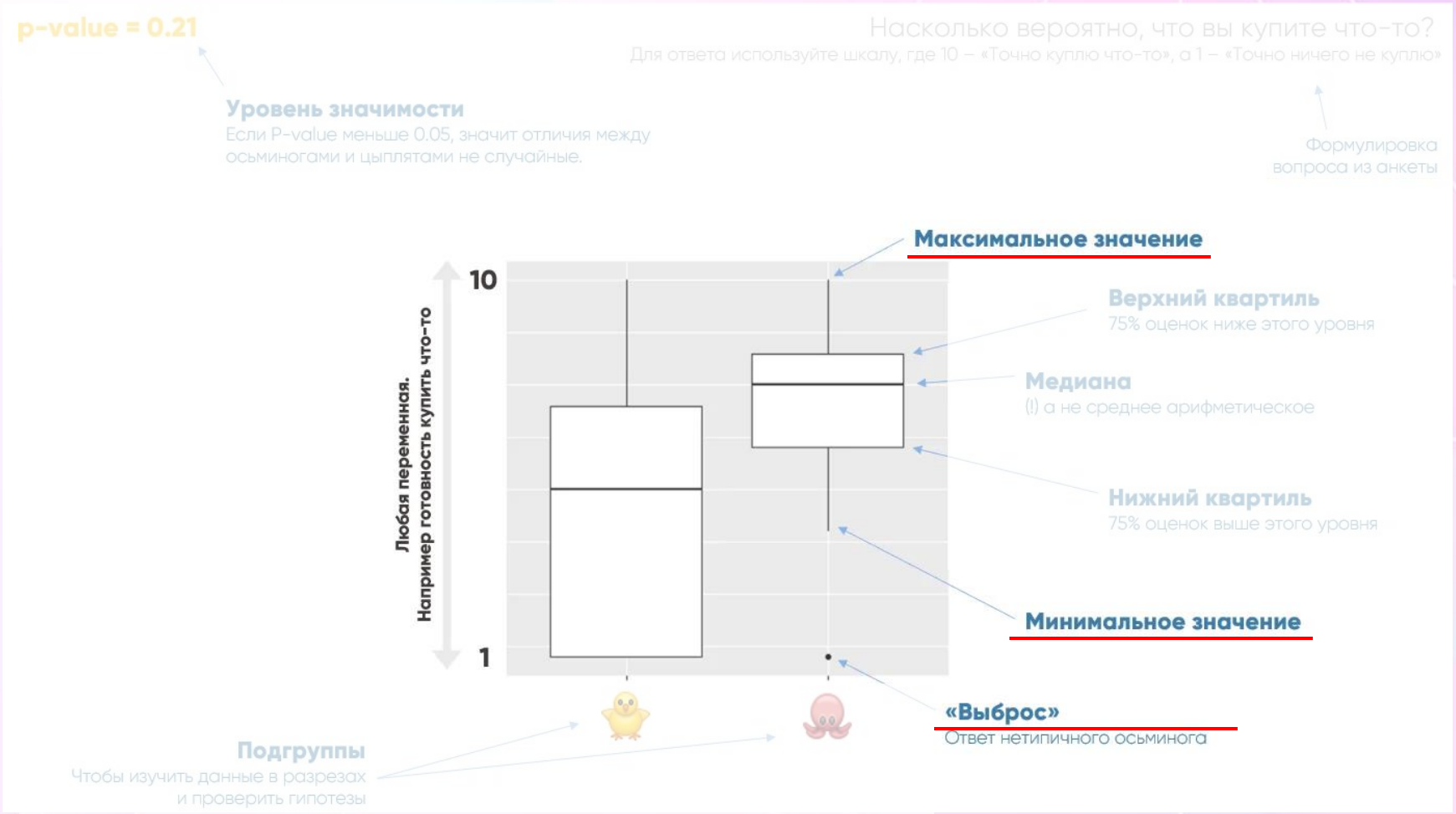
Figure C. Regular Box Plot

Здесь усы простираются до минимума и максимума в данных

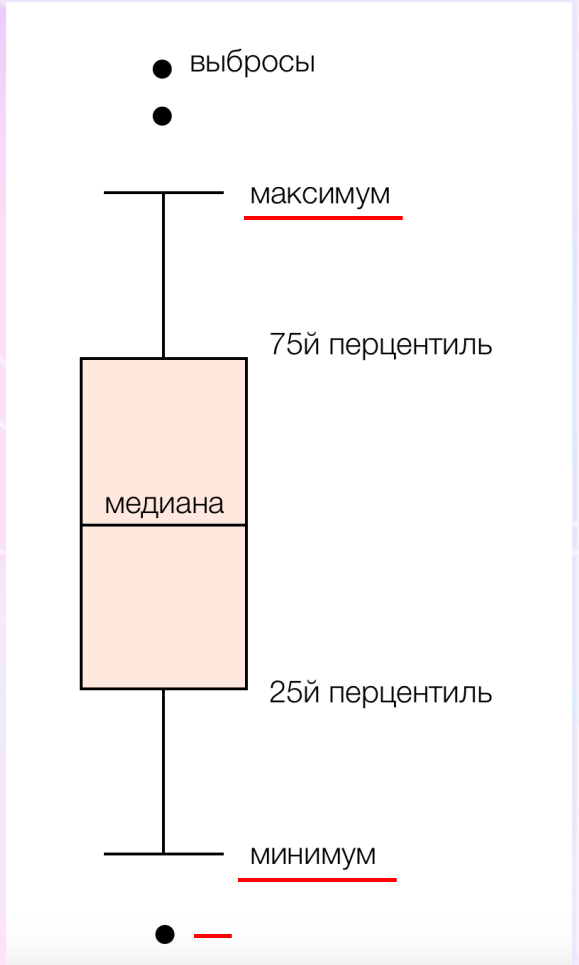
При желании можно настроить в боксплоте среднее, стандартное отклонение или стандартную ошибку

Variations of Box Plots (https://www.jstor.org/stable/2683468)

Боксплот картинки хорошие и разные



https://www.tidydata.ru/boxplot

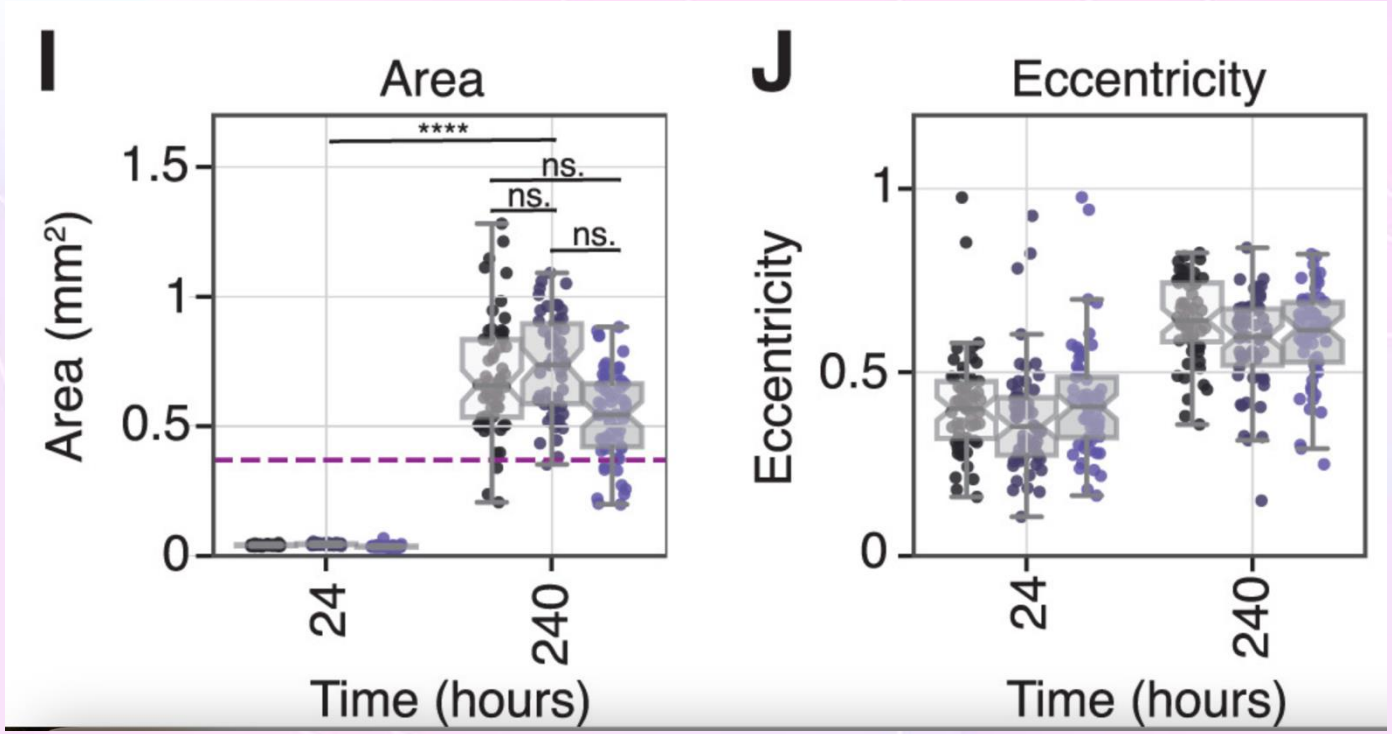


https://nastengraph.medium.com/зачем-вам-боксплоты-aa5f6d662a83

Почему-то на этих графиках усы обозначают минимум и максимум, хотя на боксплотах присутствуют выбросы

Боксплот: инструкция по использованию

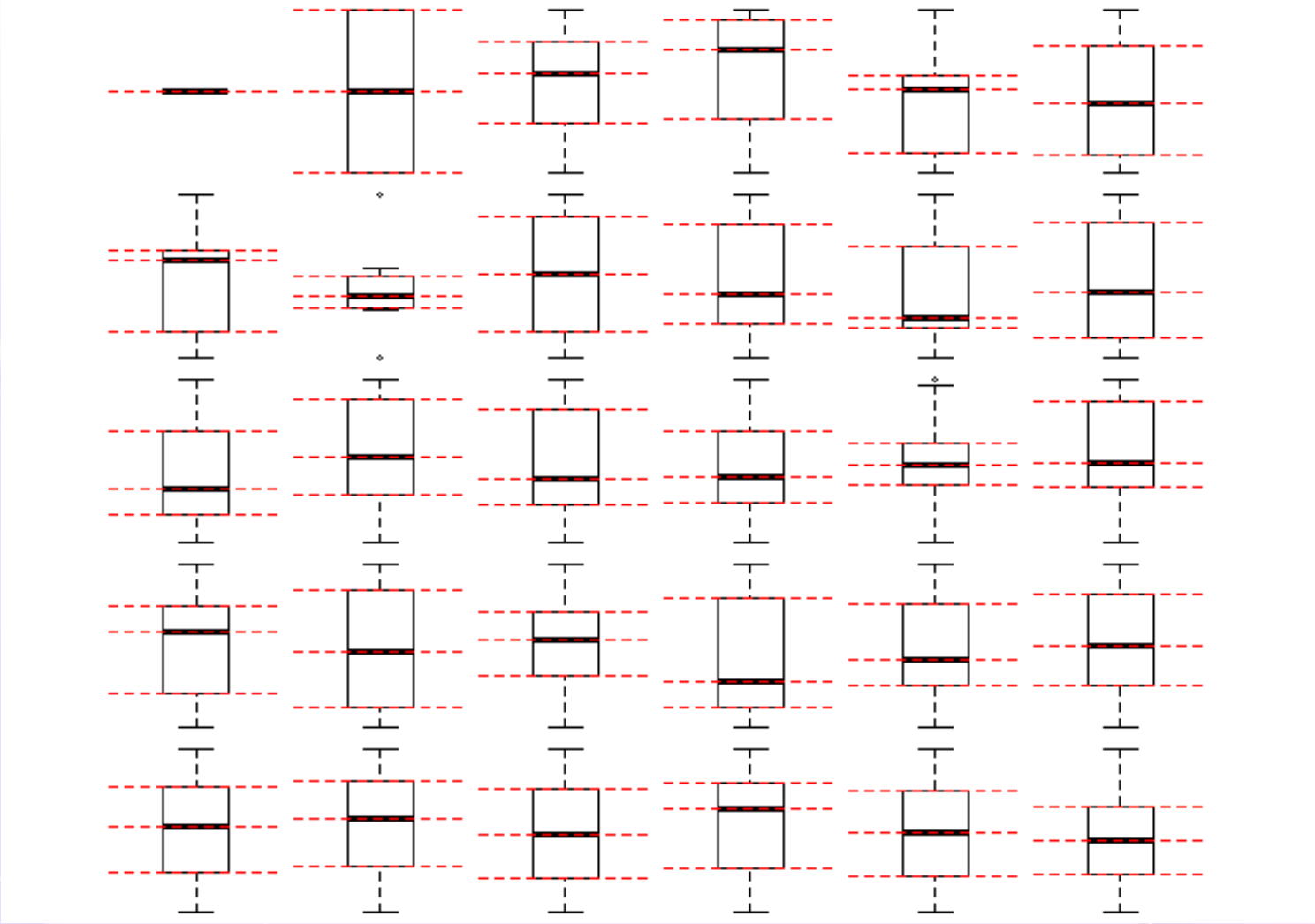
Всегда подписывайте элементы боксплота и читайте в статье описание



DOI: 10.1016/j.cell.2025.04.025

(I and J) Quantification of (I) area and (J) eccentricity for individual PGAs at days 1 and 10 for H1 hESCs. Boxplots represent median and interquartile range. **** $p < 0.00005$. Dotted line shows 0.37 mm2 cutoff. Data from 3 independent experiments are shown.

9 способов расчета квантилей



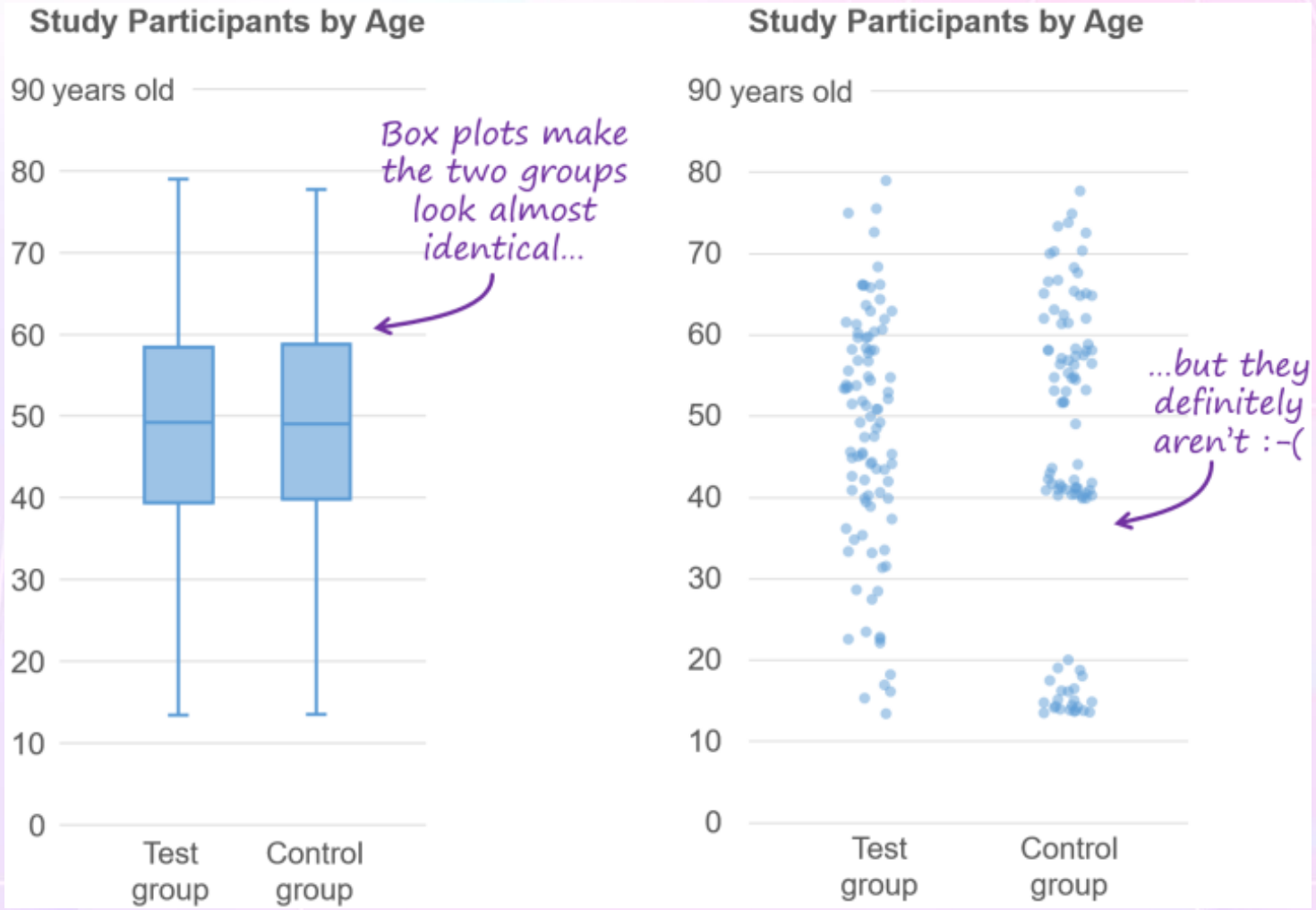
В функции `boxplot` и `summary` используются разные способы расчета квантилей по умолчанию.

Обычно они приводят к одним и тем же результатам, но в редких случаях результат может отличаться.

```
?quantile
```

https://stackoverflow.com/questions/40634693/lower-and-upper-quartiles-in-boxplot-in-r

А всегда ли нужно использовать боксплот?



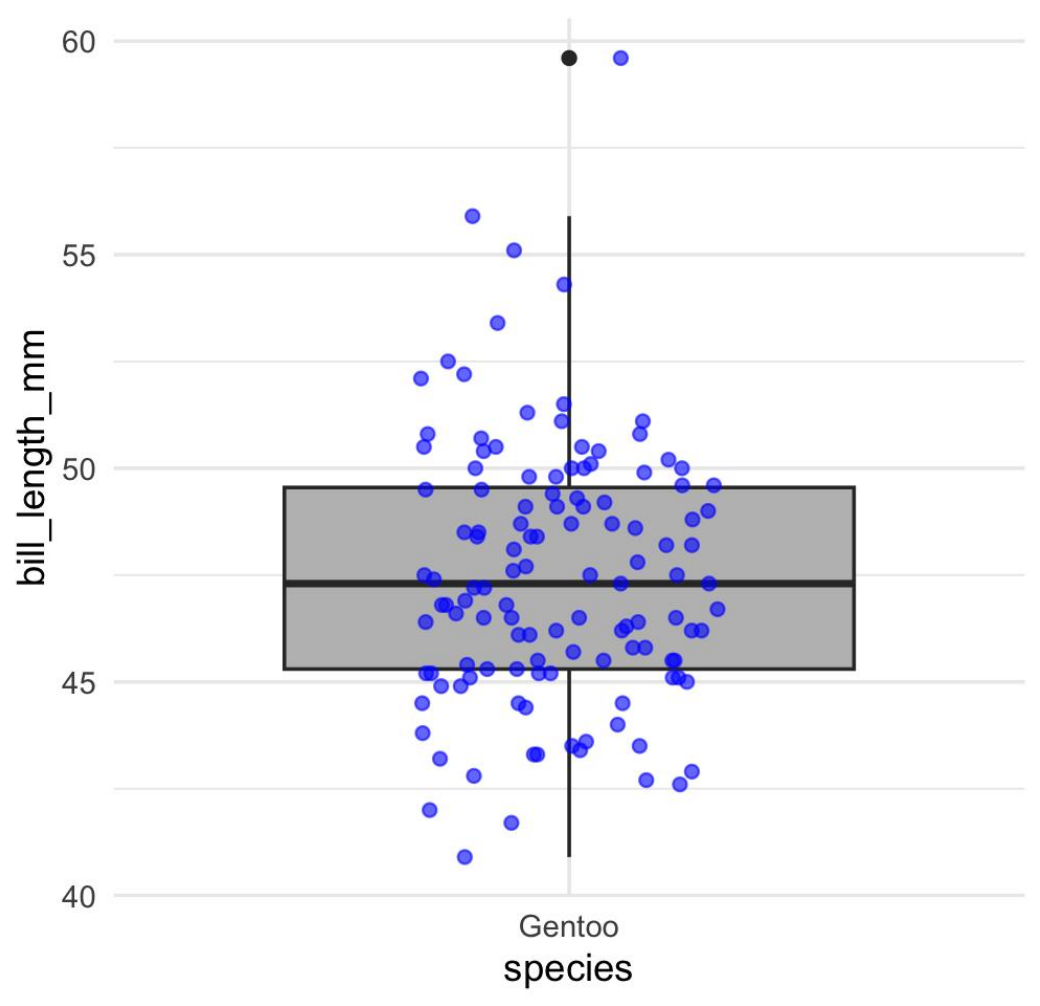
Боксплот дает мало информации о распределении данных.

Также нет информации о количестве наблюдений.

Если значений сравнительно мало, то можно их показать на графике.

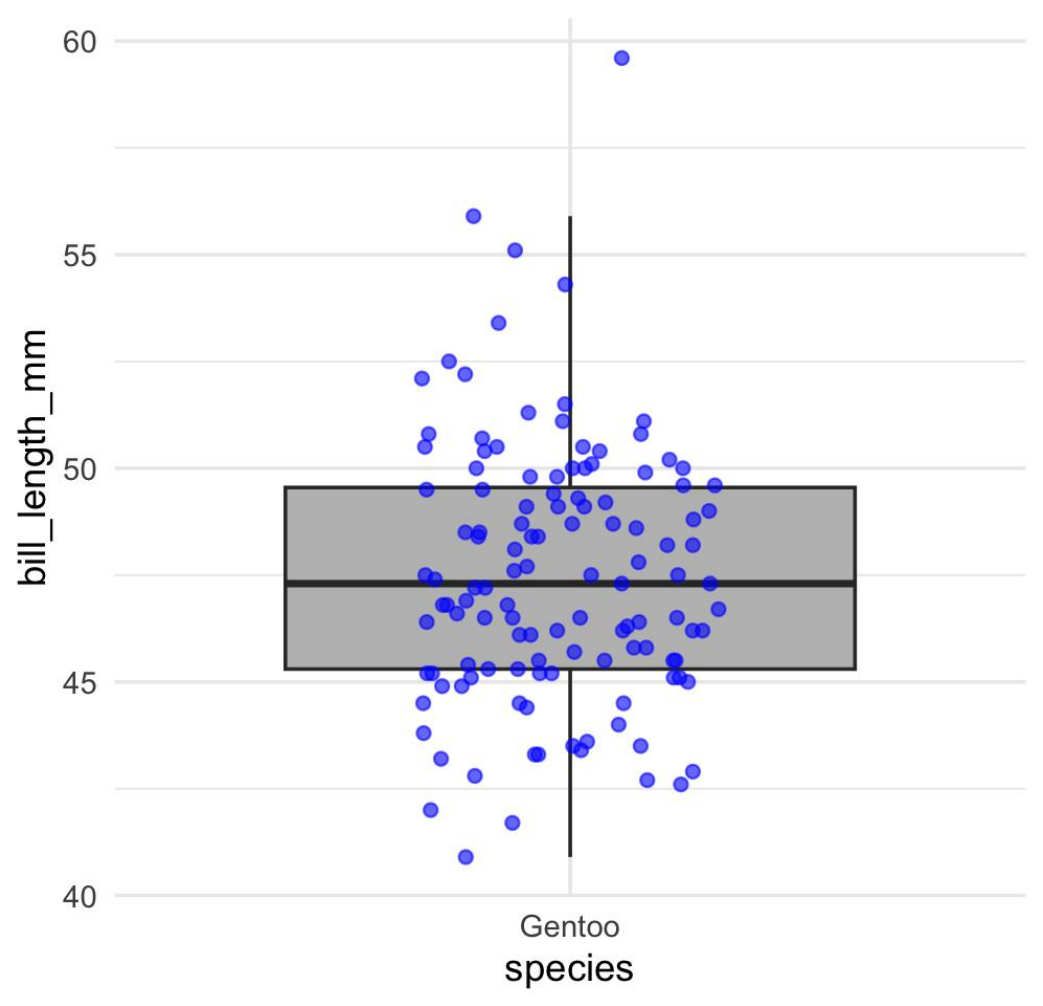
https://nightingaledvs.com/ive-stopped-using-box-plots-should-you/

Показать все наблюдения




```
set.seed(1)
penguins %>%
  filter(species == 'Gentoo') %>%
  ggplot(aes(species, bill_length_mm))+
  geom_boxplot(fill = 'gray')+
  geom_jitter(height = 0, width = 0.2, alpha
= 0.6, color = 'blue', size = 1.5)+
  theme_minimal()
```

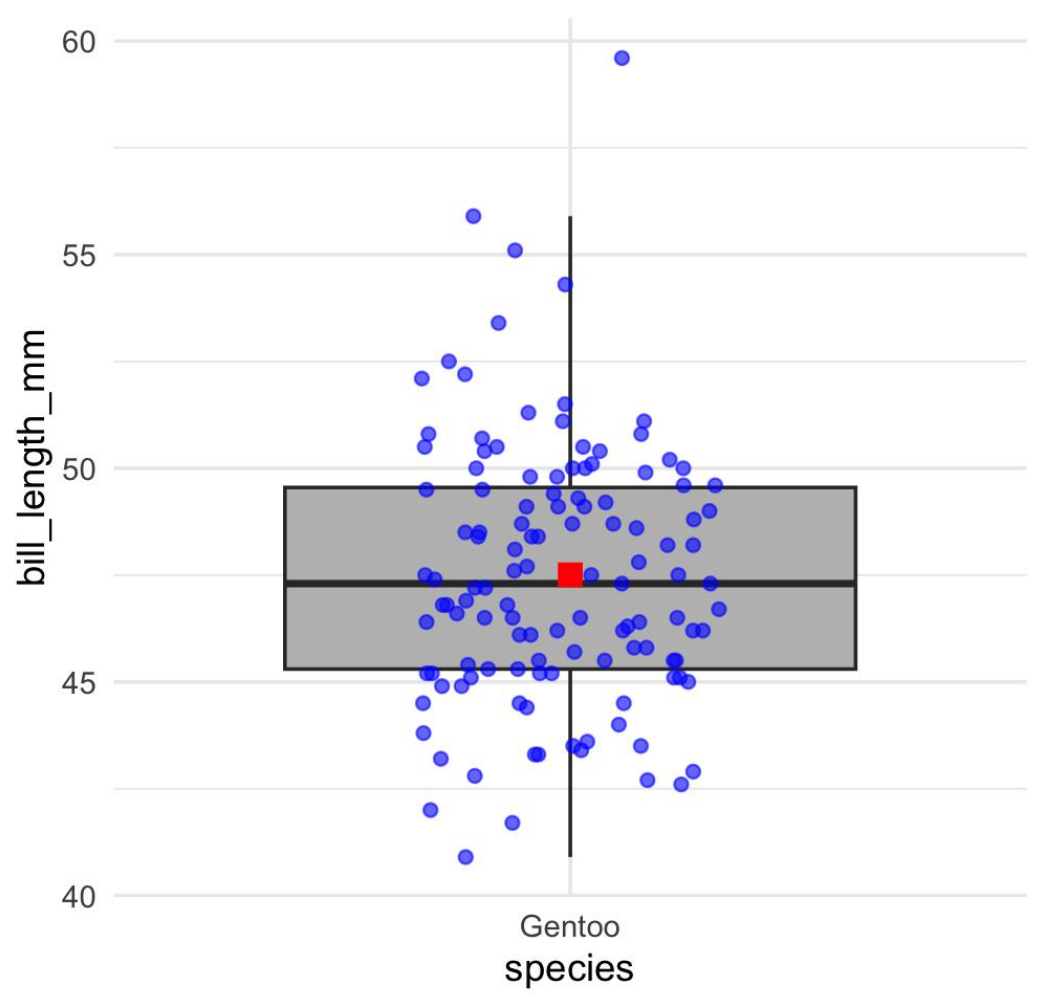
Показать все наблюдения




```
set.seed(1)
penguins %>%
  filter(species == 'Gentoo') %>%
  ggplot(aes(species, bill_length_mm))+
  geom_boxplot(fill = 'gray',
               outlier.shape = NA)+
  geom_jitter(height = 0, width = 0.2, alpha
= 0.6, color = 'blue', size = 1.5)+
  theme_minimal()
```

Убрали аутлаеры с графика, чтобы не казалось, что у нас дублированы наблюдения

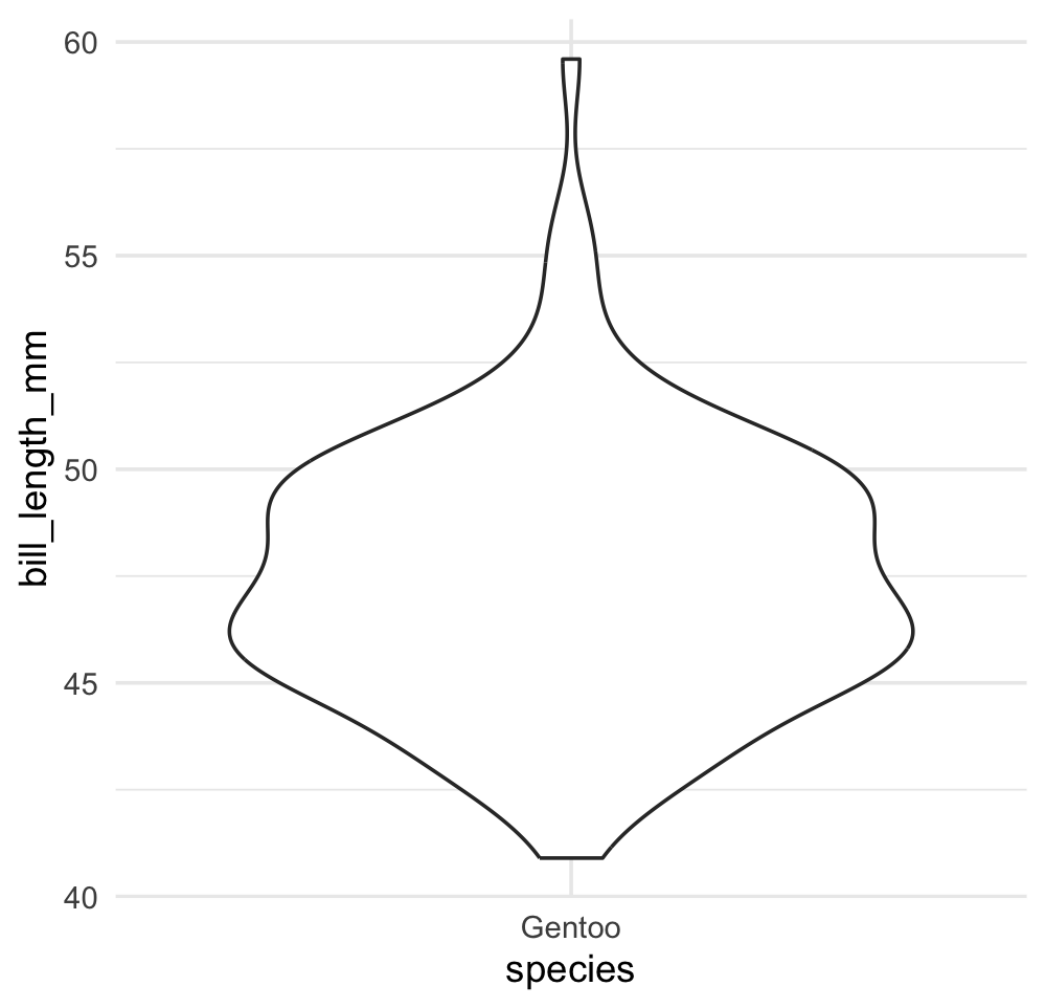
Показать все наблюдения




```
set.seed(1)
penguins %>%
  filter(species == 'Gentoo') %>%
  ggplot(aes(species, bill_length_mm))+
  geom_boxplot(fill = 'gray', outlier.shape
= NA)+
  geom_jitter(height = 0, width = 0.2, alpha
= 0.6, color = 'blue', size = 1.5)+
  stat_summary(fun.y=mean, geom="point",
shape=15, size=3, color="red", fill="red")+
  theme_minimal()
```

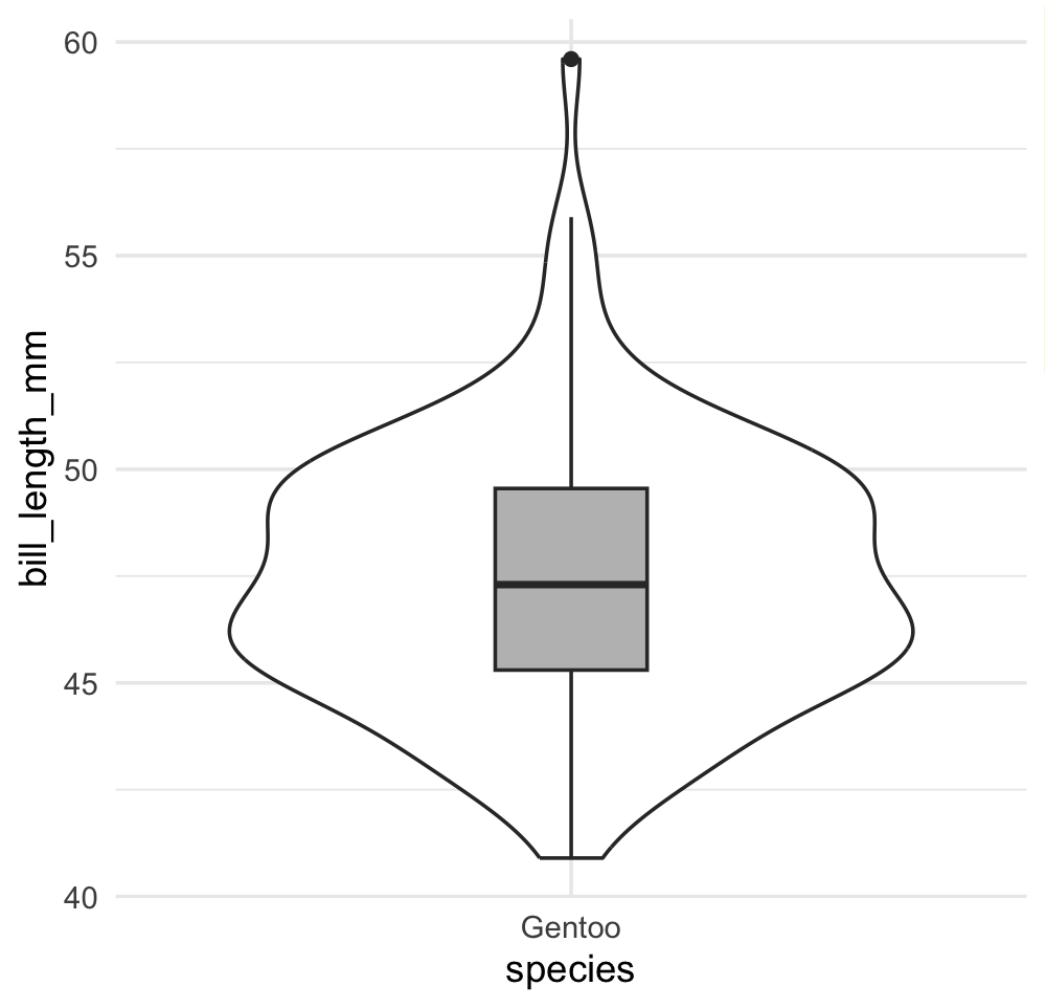
Добавили обозначение среднего на графике

А если использовать violin?



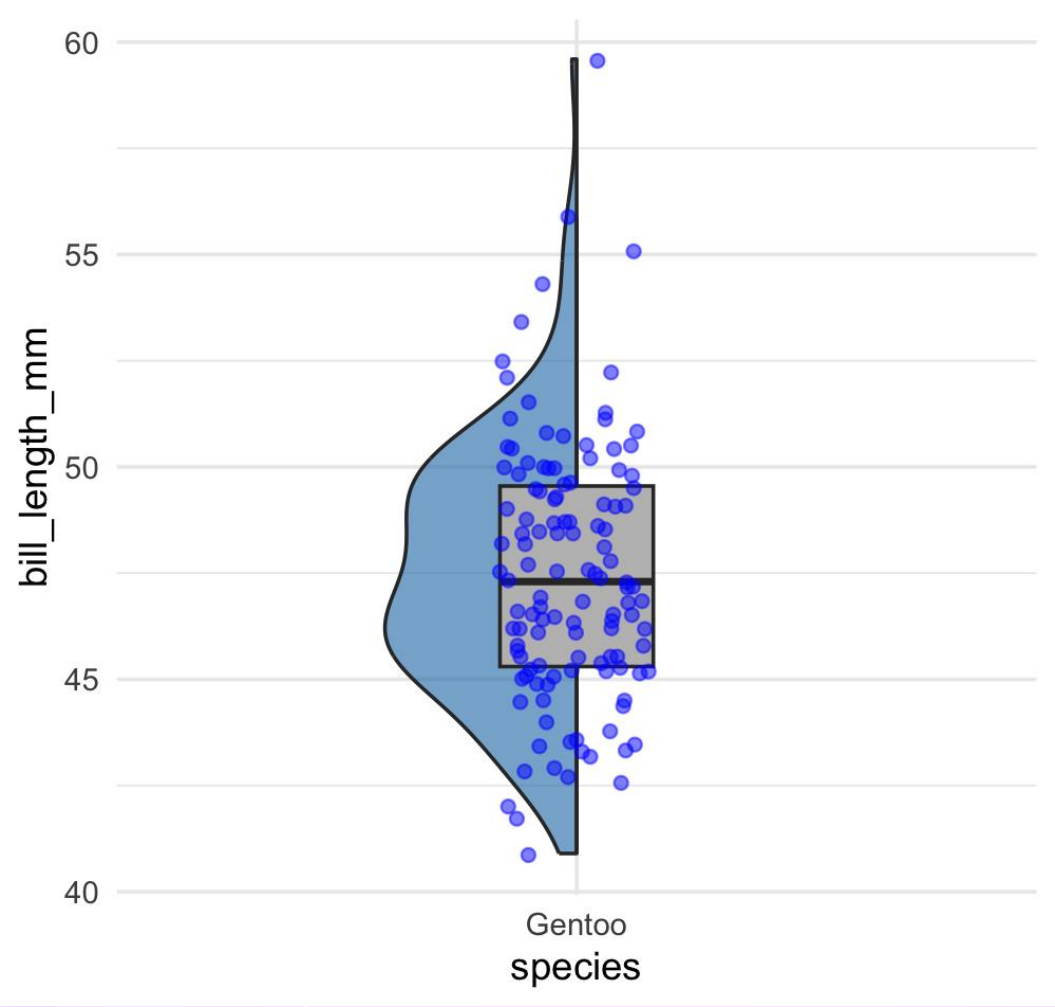

```
penguins %>%
  filter(species == 'Gentoo') %>%
  ggplot(aes(species, bill_length_mm))+
  geom_violin()+
  theme_minimal()
```

А если использовать violin + boxplot?




```
penguins %>%
  filter(species == 'Gentoo') %>%
  ggplot(aes(species, bill_length_mm))+
  geom_violin()+
  geom_boxplot(fill = 'gray', width = 0.2)+
  theme_minimal()
```

А если использовать violin + box + jitter?




```
library(gghalves)
penguins %>%
  filter(species == "Gentoo") %>%
  ggplot(aes(x = species, y = bill_length_mm))+
  geom_half_violin(side = "l", width = 0.5,
fill = "#1f77b4", alpha = 0.6)+
  geom_boxplot(fill = "gray", width = 0.2,
outlier.shape = NA  )+
  geom_jitter(width = 0.1, alpha = 0.5, color =
'blue', size = 1.5)+
  theme_minimal()
```

Интерпретация пределов погрешностей

Пределы погрешностей (error bars)

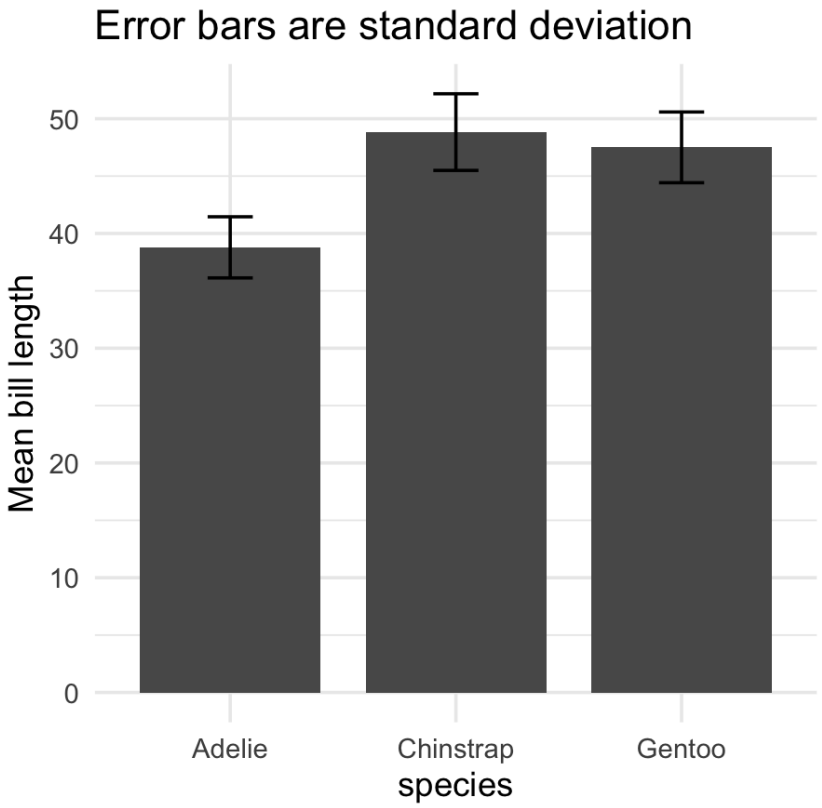
Descriptive error bars – описывают выборку

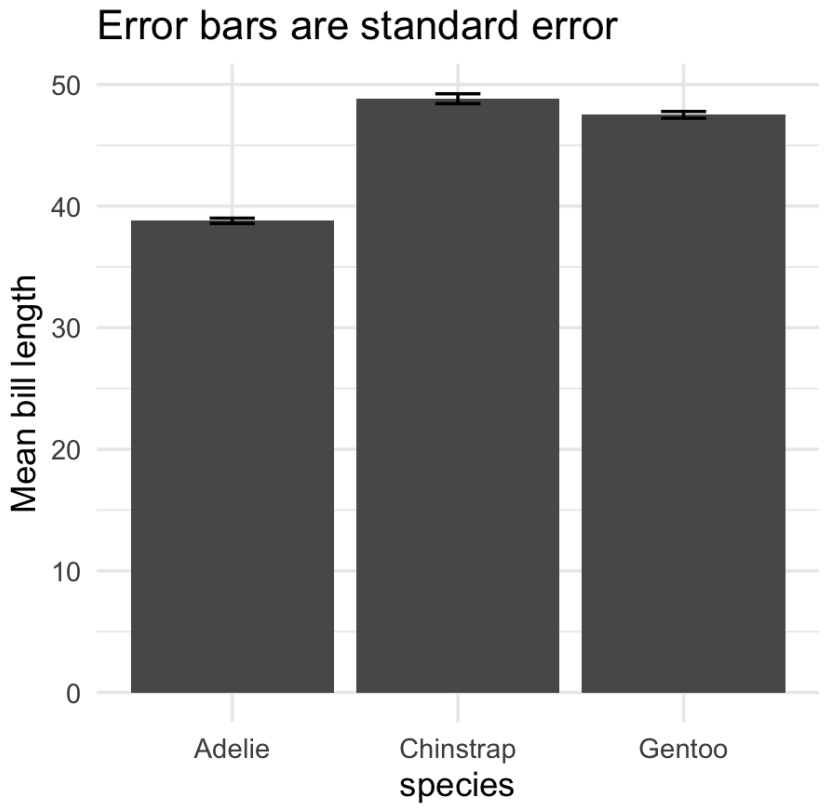
Стандартное отклонение (sd)

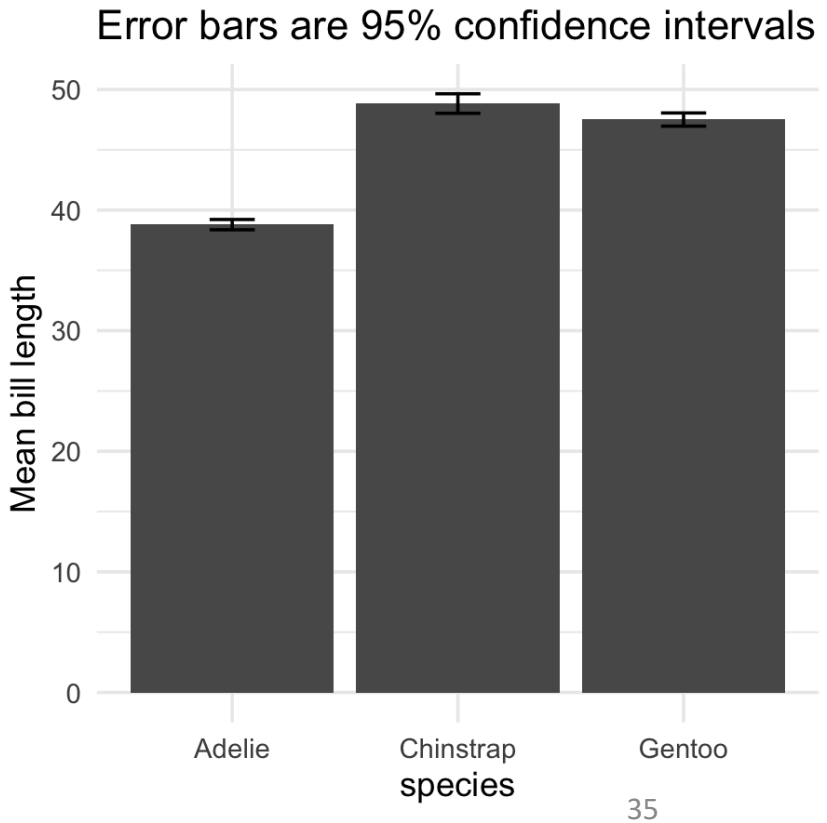
Inferential error bars – делают вывод о генеральной совокупности

Стандартная ошибка (se)

Доверительный интервал (CI)







Пределы погрешностей (error bars)

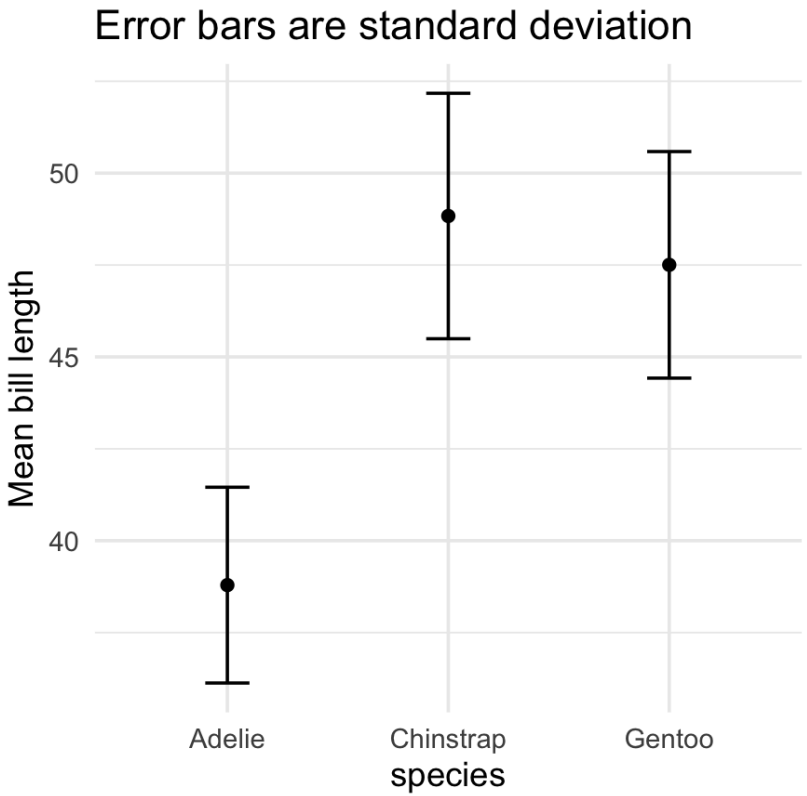
Descriptive error bars – описывают выборку

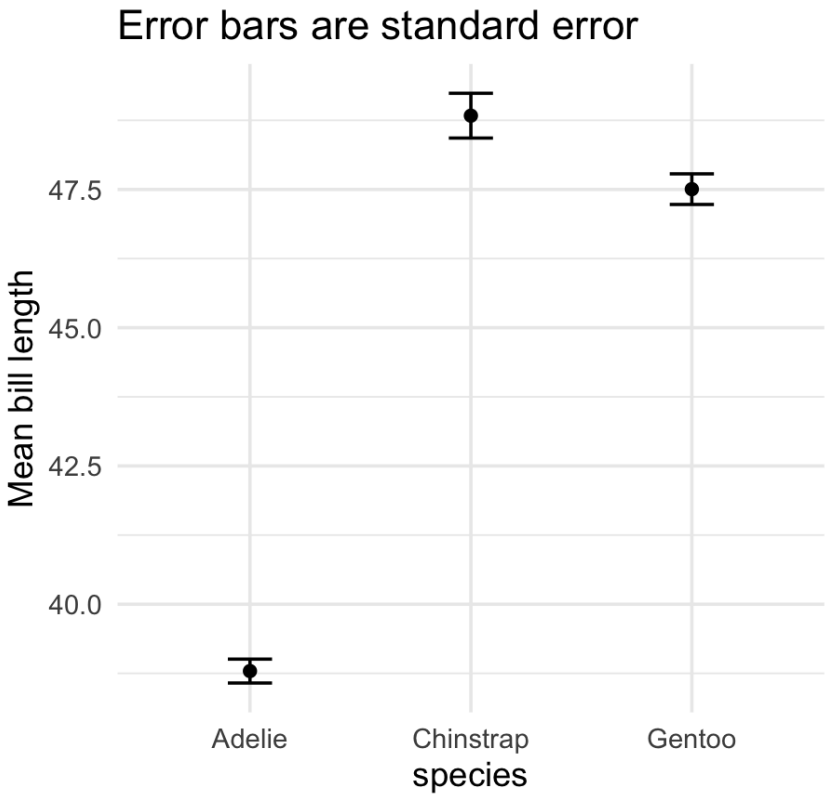
Стандартное отклонение (sd)

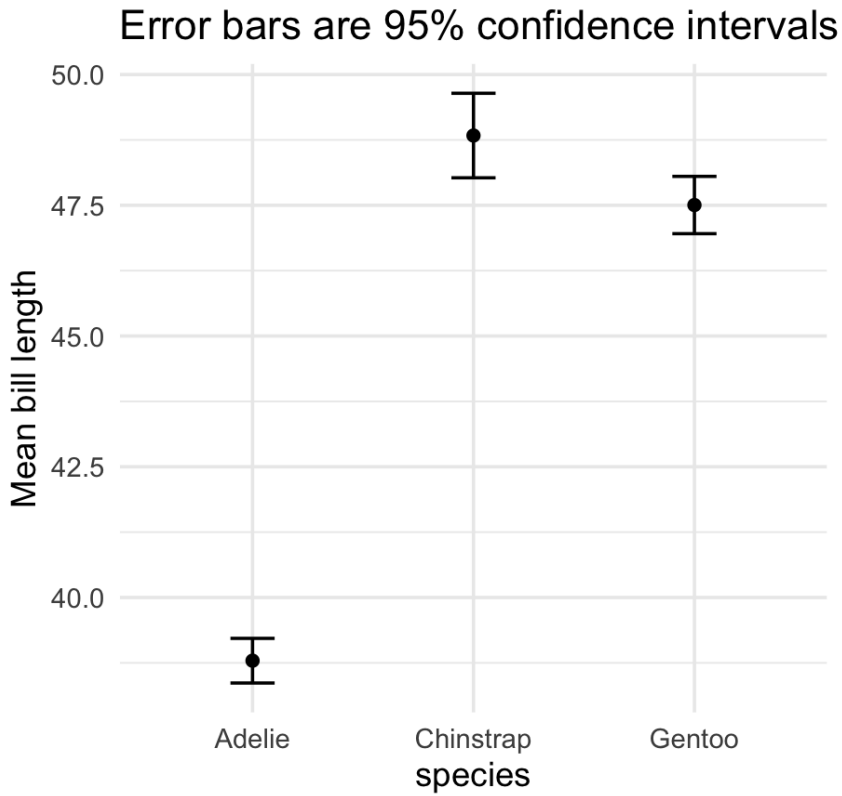
Inferential error bars – делают вывод о генеральной совокупности

Стандартная ошибка (se)

Доверительный интервал (CI)







О перекрытии и не перекрытии se и CI

Правило:

- **Перекрывание** стандартных ошибок говорит об **отсутствии** значимости различий, при этом обратное не верно
- С доверительными интервалами наоборот - **отсутствие перекрывания** 95% доверительных интервалов говорит о **значимости** различий, в то время обратное не верно.

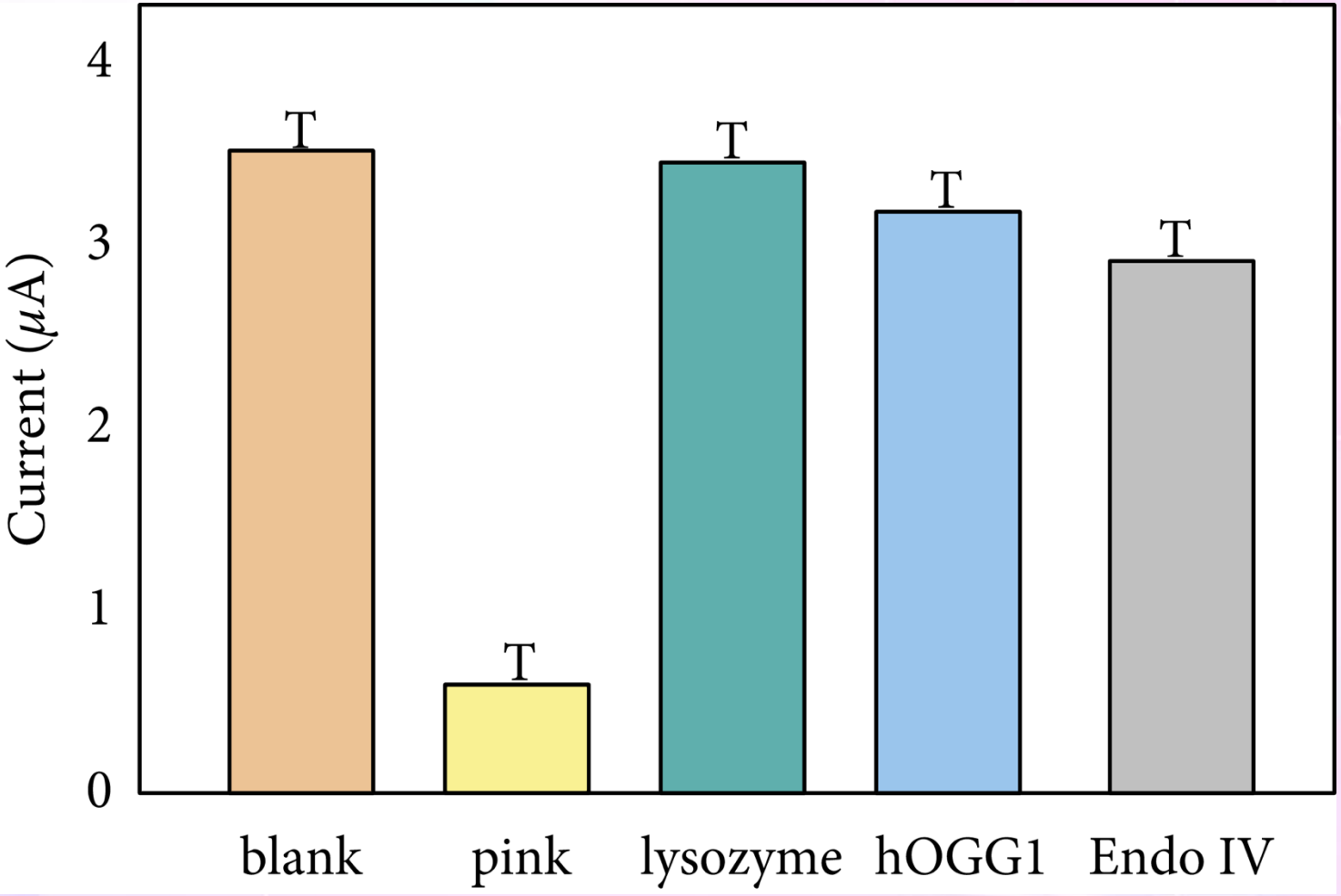
Rules of thumb (for when sample sizes are equal, or nearly equal).

Type of error bar	Conclusion if they overlap	Conclusion if they don't overlap
SD	No conclusion	No conclusion
SEM	$P > 0.05$	No conclusion
95% CI	No conclusion	$P < 0.05$ (assuming no multiple comparisons)

Но! Лучше вообще не делать выводы о статистической значимости, исходя из перекрывания пределов погрешностей, так как это еще зависит от размера выборок, поправок на множественное тестирование и тп.

https://www.graphpad.com/support/faq/spanwhat-you-can-conclude-when-two-error-bars-overlap-or-dontspan/

Как не надо делать :)



https://retractionwatch.com/2022/12/22/that-paper-with-the-t-error-bars-was-just-retracted/

Как использовать пределы погрешностей

- Обязательно подписывать, что означают пределы погрешностей на вашем графике.
- Выбирать вид пределов погрешностей, исходя из задачи.

Как надо делать графики

Excellence in statistical graphics consists of complex ideas communicated with clarity, precision, and efficiency. Graphical displays should

- show the data
- induce the viewer to think about the substance rather than about methodology, graphic design, the technology of graphic production, or something else
- avoid distorting what the data have to say
- present many numbers in a small space
- make large data sets coherent
- encourage the eye to compare different pieces of data
- reveal the data at several levels of detail, from a broad overview to the fine structure
- serve a reasonably clear purpose: description, exploration, tabulation, or decoration
- be closely integrated with the statistical and verbal descriptions of a data set.

Graphics *reveal* data. Indeed graphics can be more precise and revealing than conventional statistical computations.

Excellence in statistical graphics consists of complex ideas communicated with clarity, precision, and efficiency. Graphical displays should

- show the data
- induce the viewer to think about the substance rather than about methodology, graphic design, the technology of graphic production, or something else
- avoid distorting what the data have to say
- present many numbers in a small space
- make large data sets coherent
- encourage the eye to compare different pieces of data
- reveal the data at several levels of detail, from a broad overview to the fine structure
- serve a reasonably clear purpose: description, exploration, tabulation, or decoration
- be closely integrated with the statistical and verbal descriptions of a data set.

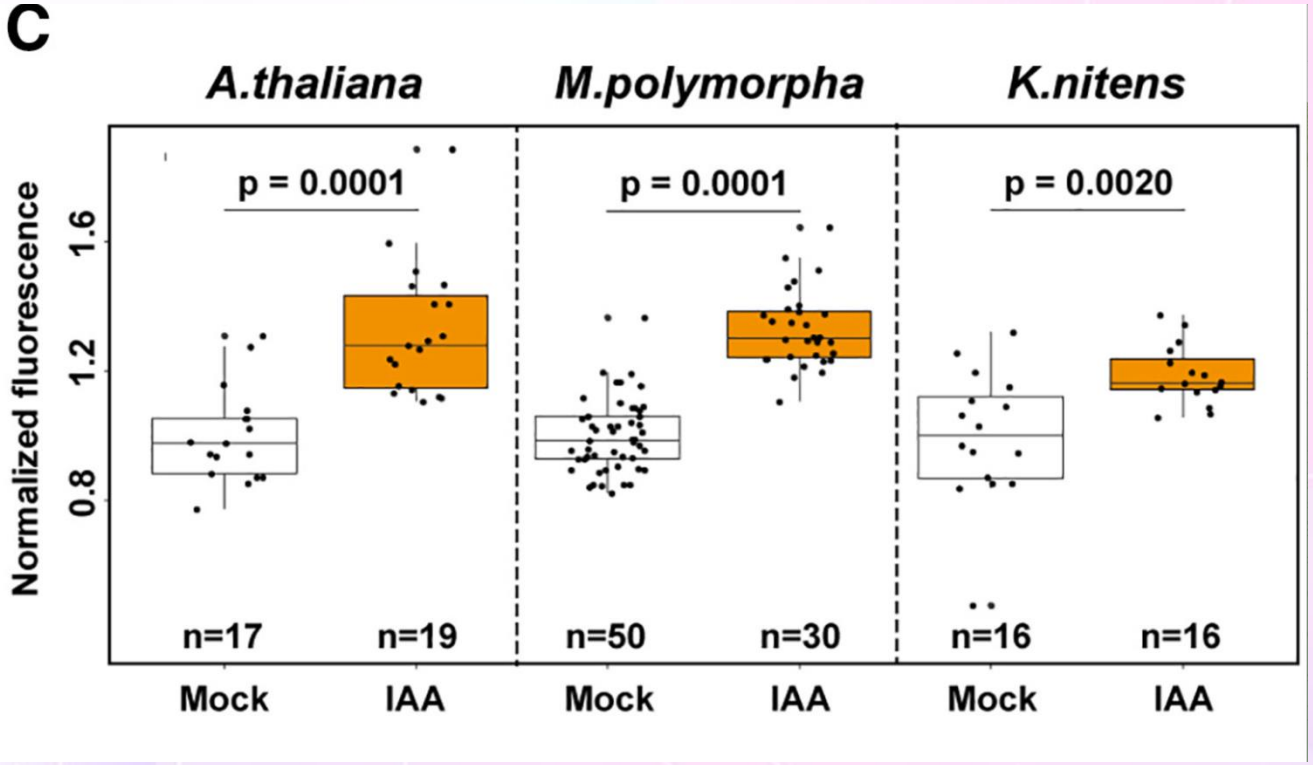
Graphics *reveal* data. Indeed graphics can be more precise and revealing than conventional statistical computations.

Хорошая визуализация:

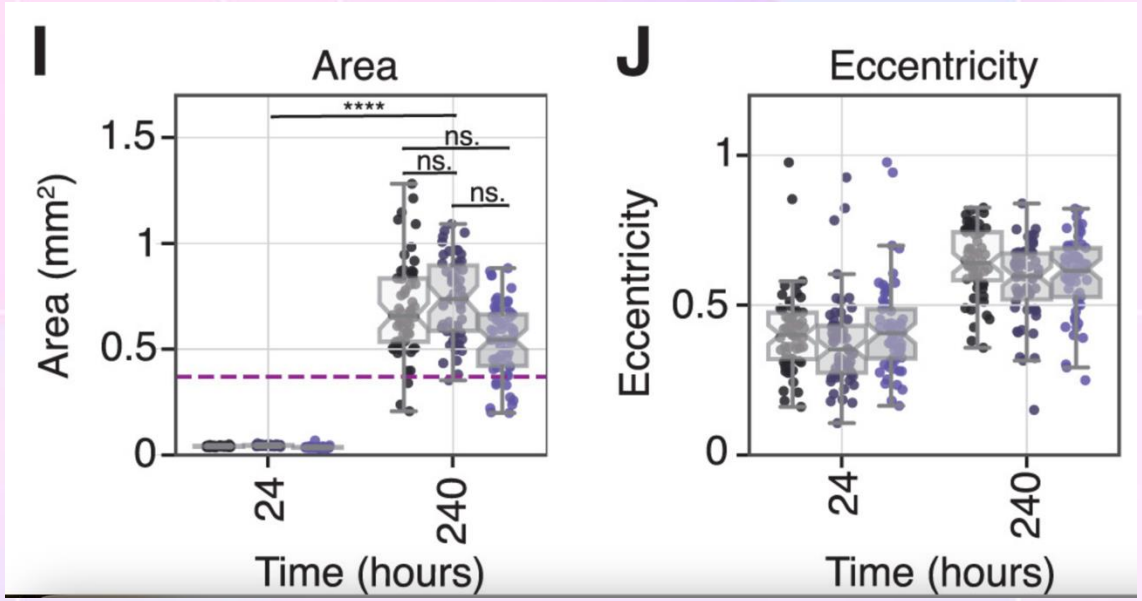
- Показывает данные
- Провоцирует пользователя думать про данные, а не про оформление, дизайн и т.п.
- Не искажает восприятие данных
- Имеет высокую плотность информации
- Делает большие наборы данных связными
- Побуждает к сравнению и выявлению особенностей в данных
- Проявляет данные на разных уровнях – от общего к частному
- Подходит для одной из целей: описания, исследования, хранения или оформления
- Тесно связано со словесным и статистическим описанием датасета

Подробно в презентации Ромы Бунина
https://revealthedata.com/Tufte.pdf

Показывает данные



DOI: 10.1016/j.cell.2023.11.021



DOI: 10.1016/j.cell.2025.04.025

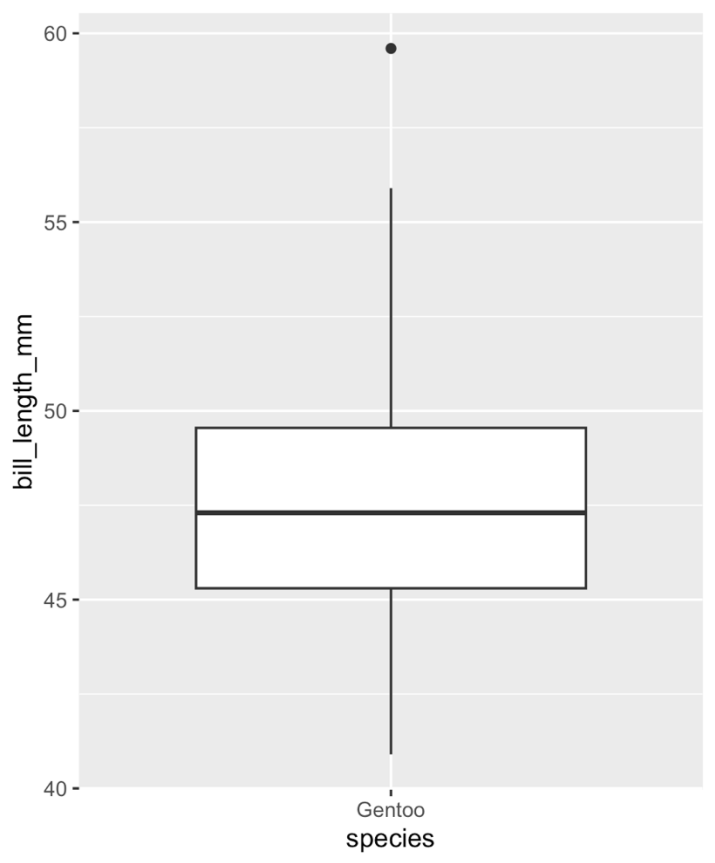
Как вы думаете, хорошая ли это визуализация?
Что можно улучшить на приведенных графиках?

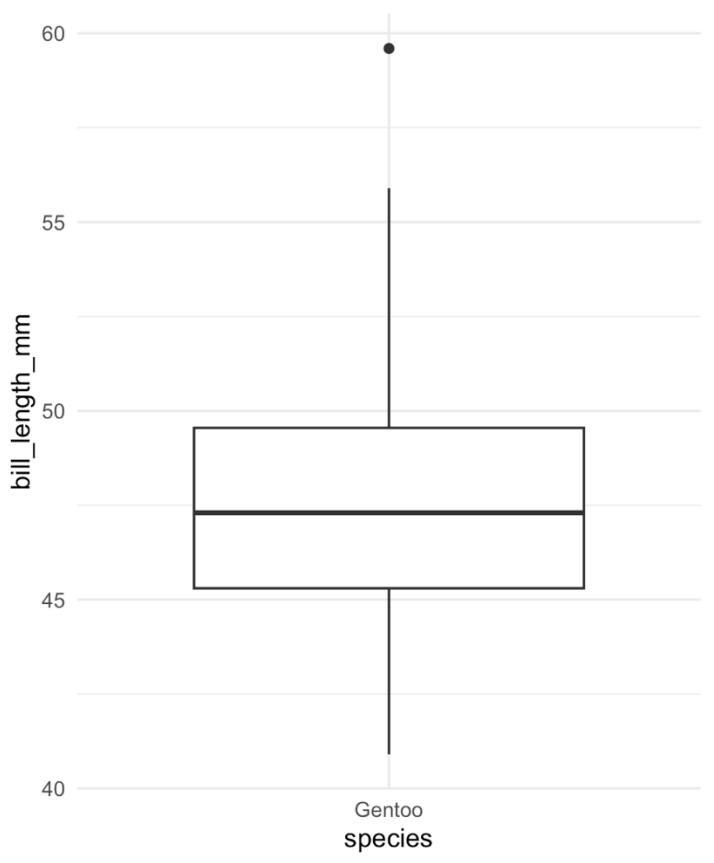
Провоцирует думать про данные

Data-ink ratio

Отношение «чернил», потраченных на данные по отношению к чернилам на всем графике

$$data\ ink\ ratio = \frac{data\ ink}{total\ ink}$$





Стандартная тема в `ggplot2` не очень хорошо соблюдает это правило

Как повысить data-ink ratio?

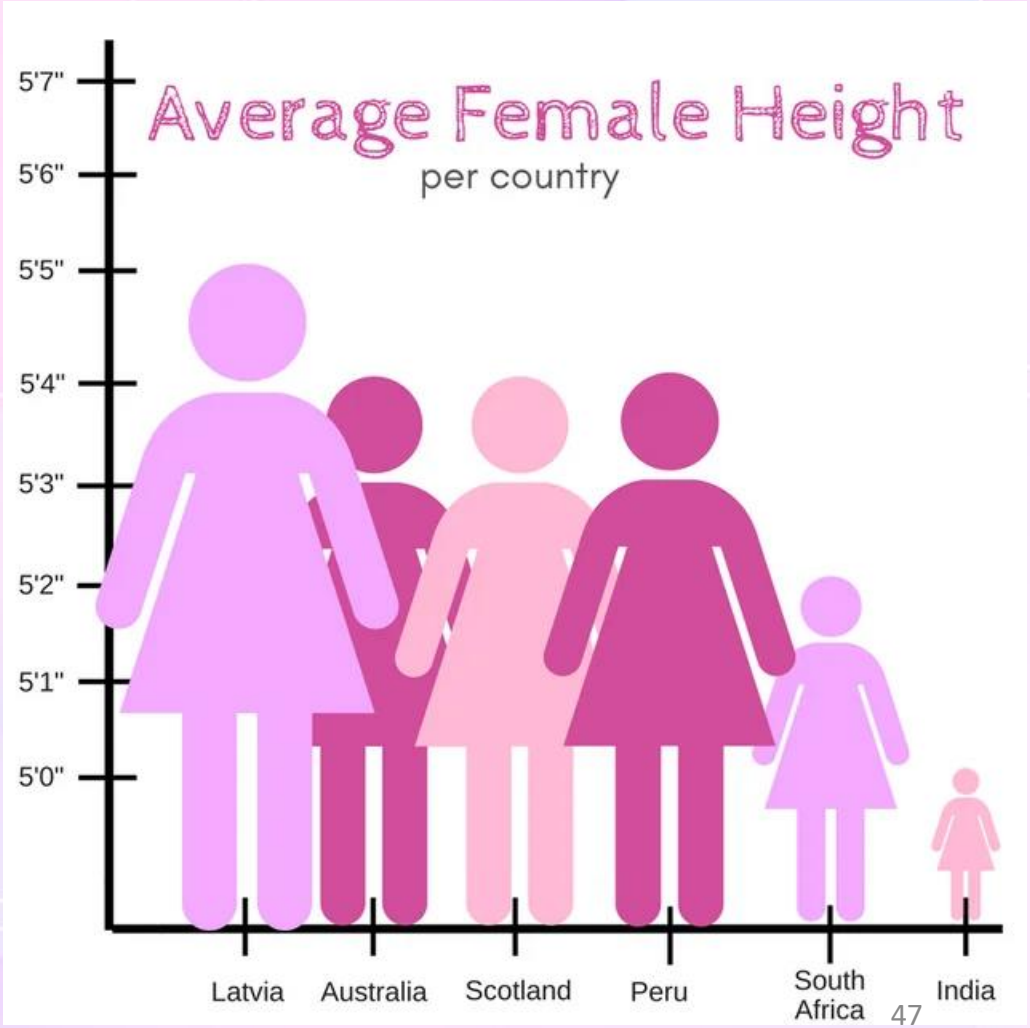
- Удалить все, что не связано напрямую с данными:
  - Удалить лишний фон
  - Удалить лишнюю сетку
  - Удалить обводку

- Удалить избыточные data-ink
  - Убрать излишнюю точность
  - Встроить легенду в график
  - Подписи вместо осей
  - Убрать лишнее цветовое кодирование

  Однако не всегда нужно бездумно следовать правилам, всегда нужно смотреть на применение в вашем случае

Не искажать восприятие данных

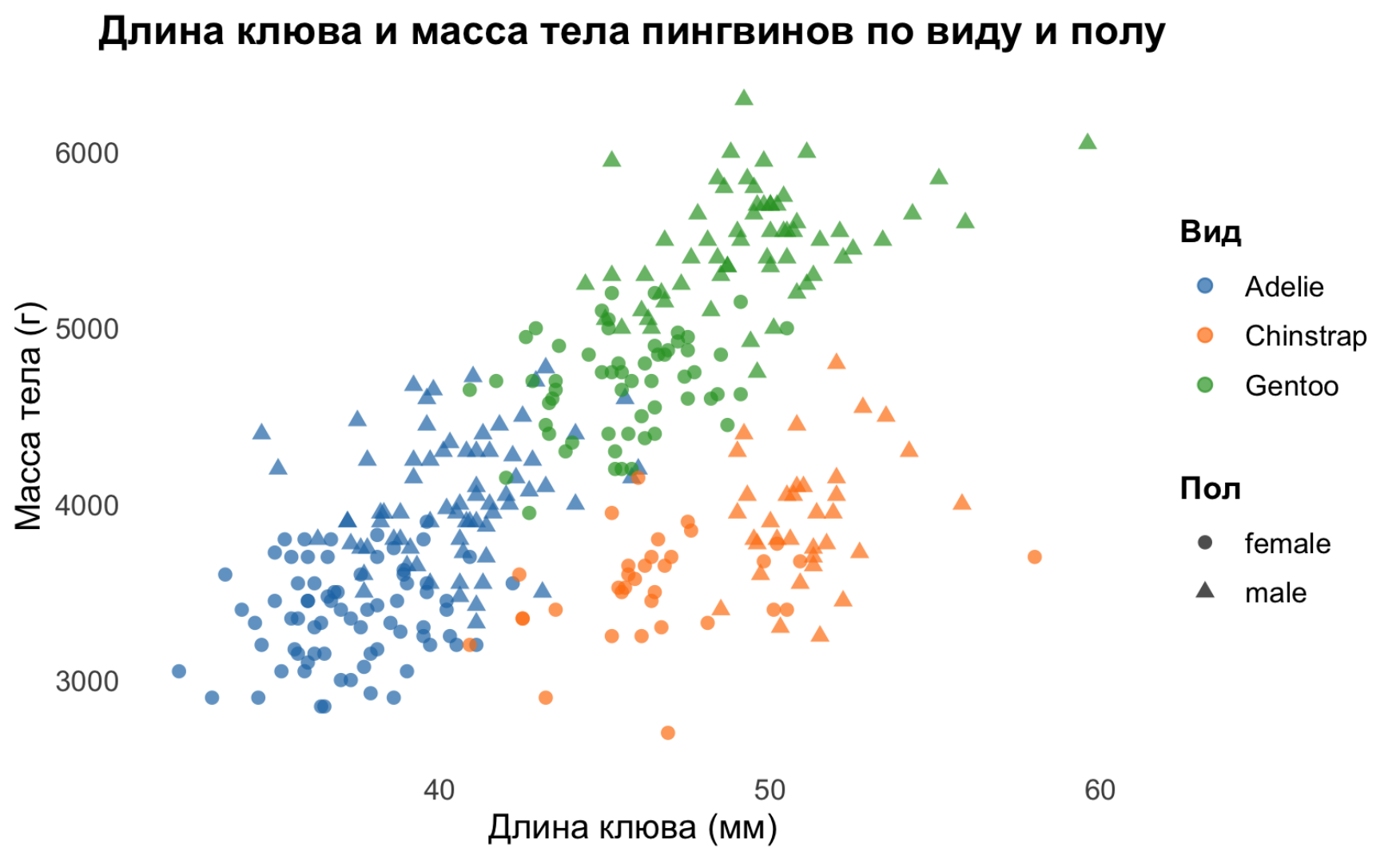
- Форма высказывания мешает понять суть
- Неверный выбор типа графика:
  - Использование пайчартов для сравнения категорий
  - Axis cropping в случае, если абсолютные значения важны
  - Спаггети-плот



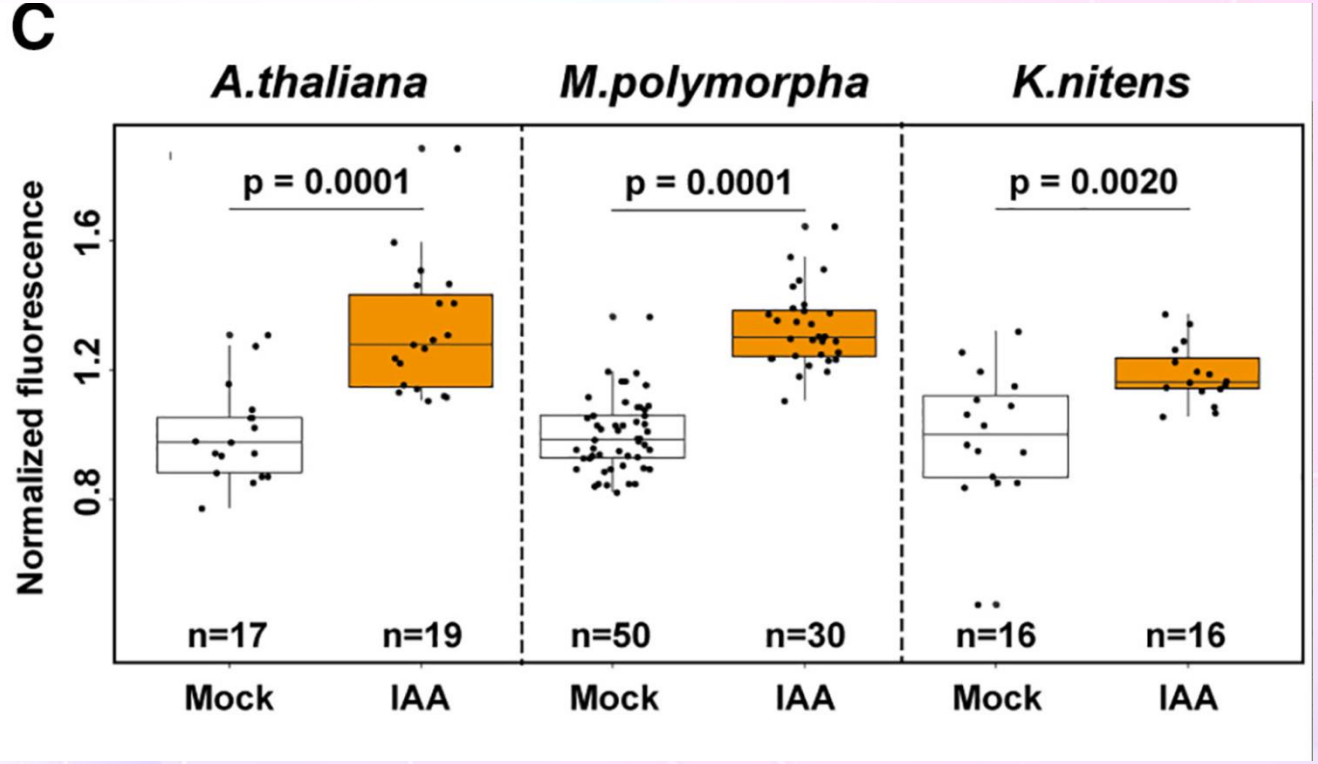
Имеет высокую плотность информации

$$\text{data density of a graphic} = \frac{\text{number of entries in data matrix}}{\text{area of data graphic}}$$

Имеет высокую плотность информации



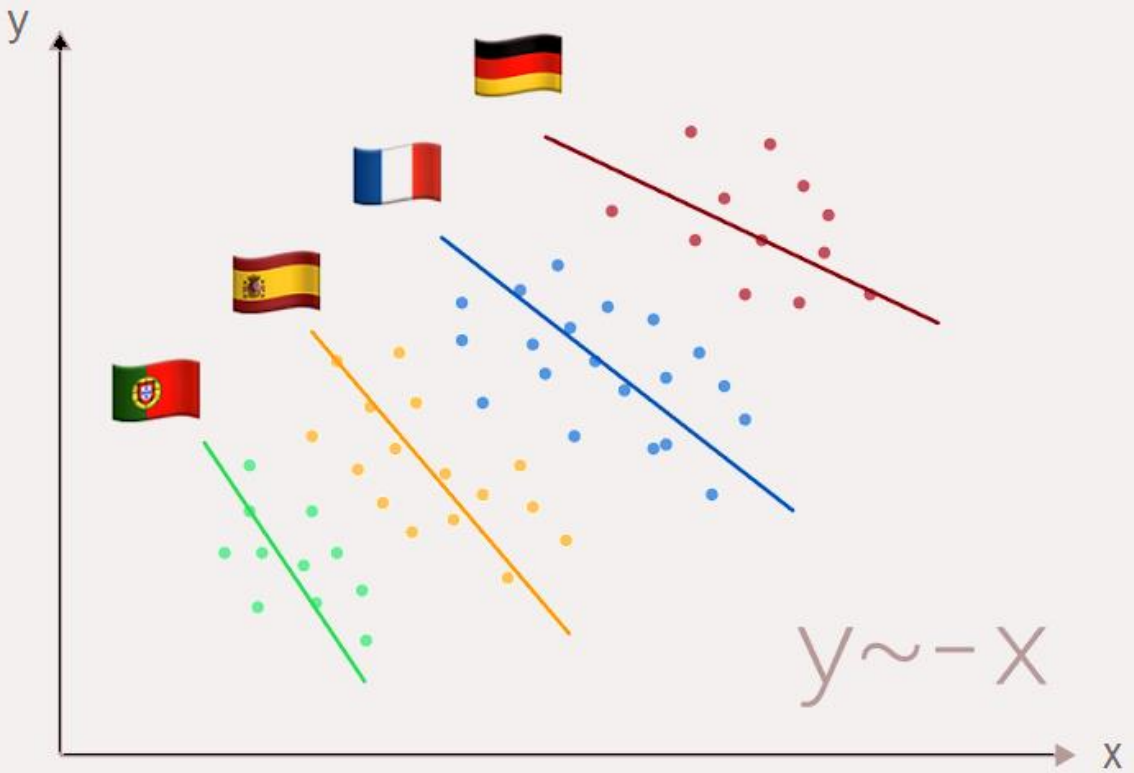
Побуждает к сравнению и выявлению особенностей в данных



DOI: 10.1016/j.cell.2023.11.021

Проявляет данные на разных уровнях – от общего к частному

Пример парадокса Сипмсона: на общем графике кажется что есть положительная корреляция между переменными, однако при разбивке на сегменты, ситуация меняется на противоположную



https://chipinfo.pro/article/paradoxes/simpsons_paradox.shtml

Тесно связано со словесным и статистическим описанием датасета

- Описание графика в тексте должно соответствовать тому, что на нем изображено.
- Сюда относится выбор типа пределов погрешностей – если задача описать данные, то стандартное отклонение, если оценка среднего, то стандартная ошибка среднего или доверительные интервалы.

Инструменты для визуализации данных

- Языки программирования
  - R: `ggplot2`
  - python: `seaborn, matplotlib, plotnine`
- GUI
  - Excel, GraphPad Prism, OriginPro
- Инструменты для BI:
  - Tableau
  - PowerBI
  - Superset
  - Datalens и тд

Ссылка на источники

- Reveal the Data https://t.me/revealthedata
- настенька и графики https://t.me/nastengraph
- душно про дату https://t.me/choking_data
- Статистика и R в науке и аналитике https://t.me/stats_for_science
- Графики лгут. Как стать информационно грамотным человеком в мире данных? | Альберто Кайро
- Fundamentals of Data Visualization | Claus O. Wilke
- The Visual Display of Quantitative Information | Edward Tufte

bioinf.institute/hack2025

https://bioinf.me/education/stat

Институт биоинформатики в социальных сетях

Разрушители статистических мифов: bioinf.me/stat_myths

Чат по биостатистике и R: https://t.me/chat_biostat_R

По всем вопросам: biostat@bioinf.me

Сайт Института: bioinf.me

Институт в VK: vk.com/bioinf

Телеграм-канал Института: t.me/bioinforussia

Чат про образование и карьеру: t.me/bioinf_career

YouTube-канал: www.youtube.com/bioinforussia