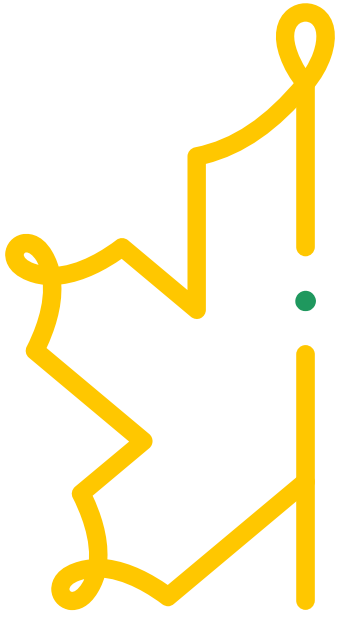
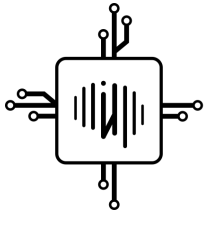


Razвед
анализ
данных

Графика

ggplot2



Фонд
интеллект

Анна Валяева

Лекция 5 - 2022

ggplot2

data =

x	y	color	size	…
16	198	cat	3	…
10	245	cat	2	…
7	178	dog	1	…
26	282	dog	2	…

mapping = aesthetic properties

x = x_col
y = y_col
color = color_col
size = size_col
…

geometric objects

geom_point()
geom_bar()
geom_boxplot()
geom_line()
…

Длинный и широкий формат

- Из широкого в длинный - `pivot longer`
- Из длинного в широкий - `pivot_wider`

wide

id	x	y	z
1	a	c	e
2	b	d	f

long

id	key	val
1	x	a
2	x	b
1	y	c
2	y	d
1	z	e
2	z	f

Широкий формат

```
relig_income
```

```
# A tibble: 18 x 11
   religion `<$10k` `$10-20k` `$20-30k` `$30-40k` `$40-50k` `$50-75k` `$75-100k`
   <chr>      <dbl>     <dbl>     <dbl>     <dbl>     <dbl>     <dbl>      <dbl>
 1 Agnostic      27        34        60        81        76       137        122
 2 Atheist       12        27        37        52        35        70         73
 3 Buddhist      27        21        30        34        33        58         62
 4 Catholic     418       617       732       670       638      1116        949
 5 Don't k~      15        14        15        11        10        35         21
 6 Evangel~     575       869      1064       982       881      1486        949
 7 Hindu          1         9         7         9        11        34         47
 8 Histori~     228       244       236       238       197       223        131
 9 Jehovah~      20        27        24        24        21        30         15
10 Jewish        19        19        25        25        30        95         69
11 Mainlin~     289       495       619       655       651      1107        939
12 Mormon        29        40        48        51        56       112         85
13 Muslim         6         7         9        10         9        23         16
14 Orthodox      13        17        23        32        32        47         38
15 Other C~       9         7        11        13        13        14         18
16 Other F~      20        33        40        46        49        63         46
17 Other W~       5         2         3         4         2         7          3
18 Unaffil~     217       299       374       365       341       528        407
# ... with 3 more variables: $100-150k <dbl>, >150k <dbl>,
#   Don't know/refused <dbl>
```

Pivot_longer

Для построения графиков с **{ggplot2}**, работы с группами категорий.

```
relig_income %>%
  pivot_longer(!religion, names_to = "income", values_to = "count")
```

```
# A tibble: 180 x 3
   religion income             count
   <chr>    <chr>              <dbl>
 1 Agnostic <$10k                27
 2 Agnostic $10-20k              34
 3 Agnostic $20-30k              60
 4 Agnostic $30-40k              81
 5 Agnostic $40-50k              76
 6 Agnostic $50-75k             137
 7 Agnostic $75-100k            122
 8 Agnostic $100-150k           109
 9 Agnostic >150k                84
10 Agnostic Don't know/refused   96
# ... with 170 more rows
```

Длинный формат

```
fish_encounters
```

```
# A tibble: 114 x 3
   fish  station  seen
   <fct> <fct>   <int>
 1 4842  Release     1
 2 4842  I80_1       1
 3 4842  Lisbon      1
 4 4842  Rstr        1
 5 4842  Base_TD     1
 6 4842  BCE         1
 7 4842  BCW         1
 8 4842  BCE2        1
 9 4842  BCW2        1
10 4842  MAE         1
# ... with 104 more rows
```

Pivot_wider

Для построения тепловых карт.

```
fish_encounters %>%
  pivot_wider(names_from = fish, values_from = seen)
```

```
# A tibble: 11 x 20
   station `4842` `4843` `4844` `4845` `4847` `4848` `4849` `4850` `4851` `4854`
   <fct>    <int>  <int>  <int>  <int>  <int>  <int>  <int>  <int>  <int>  <int>
 1 Release      1      1      1      1      1      1      1      1      1      1
 2 I80_1        1      1      1      1      1      1      1      1      1      1
 3 Lisbon       1      1      1      1      1      1     NA     NA     NA     NA
 4 Rstr         1      1      1      1     NA      1     NA      1     NA     NA
 5 Base_TD      1      1      1      1     NA     NA     NA      1     NA     NA
 6 BCE          1      1      1     NA     NA     NA     NA      1     NA     NA
 7 BCW          1      1      1     NA     NA     NA     NA      1     NA     NA
 8 BCE2         1      1      1     NA     NA     NA     NA     NA     NA     NA
 9 BCW2         1      1      1     NA     NA     NA     NA     NA     NA     NA
10 MAE          1      1      1     NA     NA     NA     NA     NA     NA     NA
11 MAW          1      1      1     NA     NA     NA     NA     NA     NA     NA
# ... with 9 more variables: 4855 <int>, 4857 <int>, 4858 <int>, 4859 <int>,
#   4861 <int>, 4862 <int>, 4863 <int>, 4864 <int>, 4865 <int>
```

Датасет про пингвинов

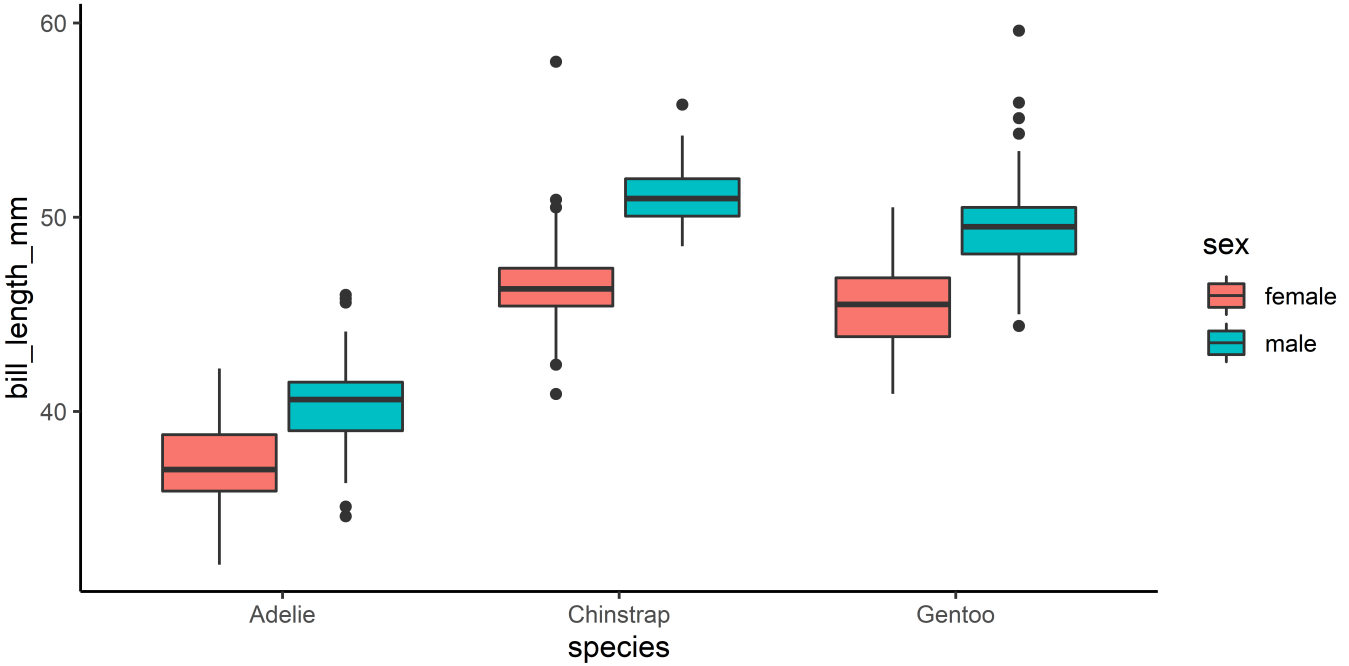
Из пакета **palmerpenguins**.

```
penguins <- read_csv("data/lesson-04/penguins.csv")
penguins
```

```
# A tibble: 344 x 8
   species island    bill_length_mm bill_depth_mm flipper_length_mm body_mass_g
   <chr>   <chr>              <dbl>         <dbl>             <dbl>       <dbl>
 1 Adelie  Torgersen           39.1          18.7               181        3750
 2 Adelie  Torgersen           39.5          17.4               186        3800
 3 Adelie  Torgersen           40.3          18                 195        3250
 4 Adelie  Torgersen           NA            NA                  NA          NA
 5 Adelie  Torgersen           36.7          19.3               193        3450
 6 Adelie  Torgersen           39.3          20.6               190        3650
 7 Adelie  Torgersen           38.9          17.8               181        3625
 8 Adelie  Torgersen           39.2          19.6               195        4675
 9 Adelie  Torgersen           34.1          18.1               193        3475
10 Adelie  Torgersen           42            20.2               190        4250
# ... with 334 more rows, and 2 more variables: sex <chr>, year <dbl>
```

Длинный формат

```
drop_na(penguins, sex) %>% ggplot() +
  geom_boxplot(aes(x = species, y = bill_length_mm, fill = sex))
```

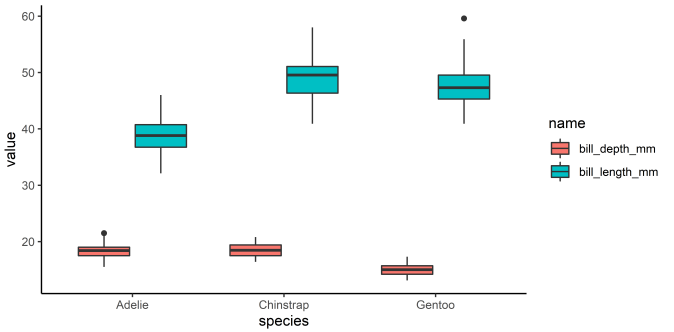


Еще длиннее

```
penguins %>%
  pivot_longer(cols = c(bill_length_mm,
    bill_depth_mm)) %>%
  select(species, name, value)
```

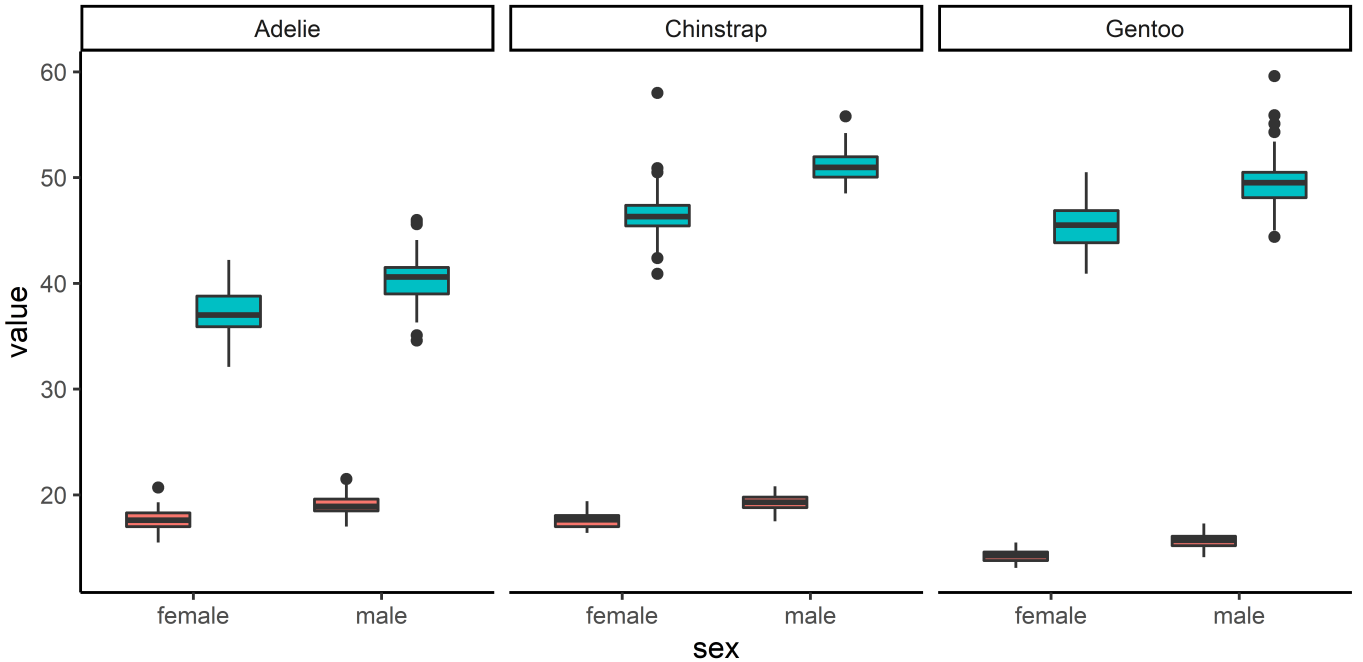
```
# A tibble: 688 x 3
   species name           value
   <chr>   <chr>          <dbl>
 1 Adelie  bill_length_mm  39.1
 2 Adelie  bill_depth_mm   18.7
 3 Adelie  bill_length_mm  39.5
 4 Adelie  bill_depth_mm   17.4
 5 Adelie  bill_length_mm  40.3
 6 Adelie  bill_depth_mm   18
 7 Adelie  bill_length_mm  NA
 8 Adelie  bill_depth_mm   NA
 9 Adelie  bill_length_mm  36.7
10 Adelie  bill_depth_mm   19.3
# ... with 678 more rows
```

```
penguins %>%
  pivot_longer(cols = c(bill_length_mm,
    bill_depth_mm)) %>%
  ggplot() +
  geom_boxplot(aes(x = species, y = value, fill =
    name))
```



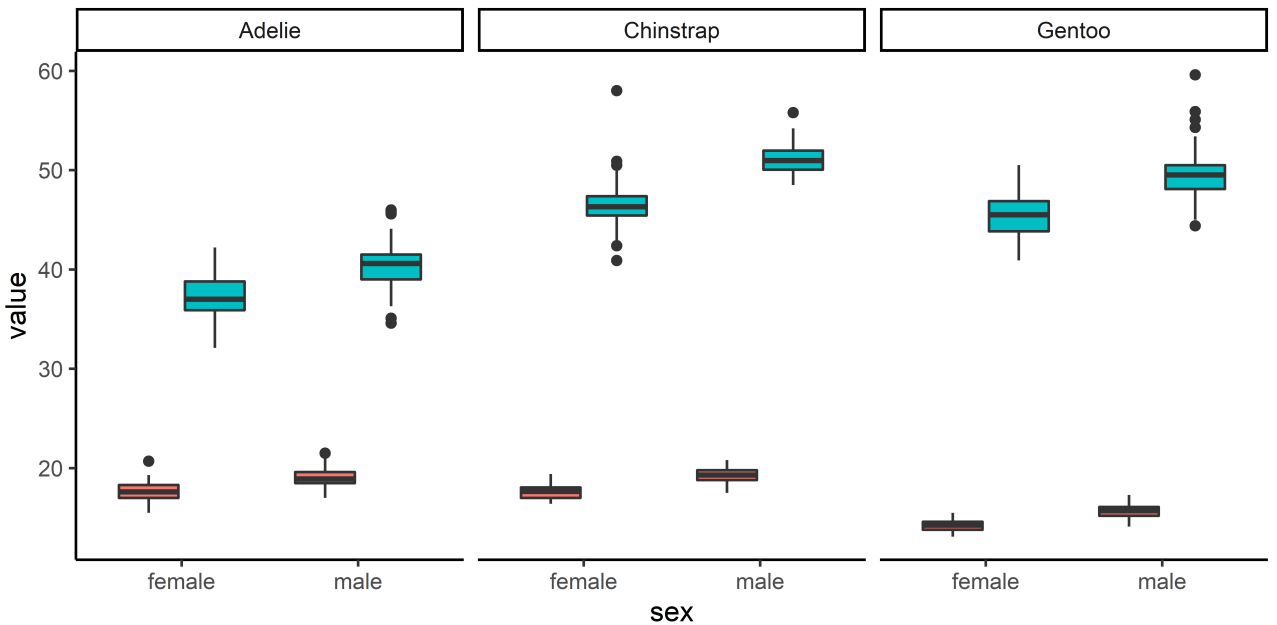
Facets

```
drop_na(penguins, sex) %>%
  pivot_longer(cols = c(bill_length_mm, bill_depth_mm)) %>%
  ggplot() +
  geom_boxplot(aes(x = sex, y = value, fill = name)) +
  facet_wrap(~ species) +
  theme(legend.position = "none")
```



Facet's labels

```
drop_na(penguins, sex) %>%
  pivot_longer(cols = c(bill_length_mm, bill_depth_mm)) %>%
  ggplot() +
  geom_boxplot(aes(x = sex, y = value, fill = name)) +
  facet_wrap(~ species,
    labeller = as_labeller(list("Адели" = "Adelie",
                                "Антарктический" = "Chinstrap",
                                "Субантарктический" = "Gentoo"))) +
  theme(legend.position = "none")
```



Комбинирование графиков

Patchwork manual Cowplot manual

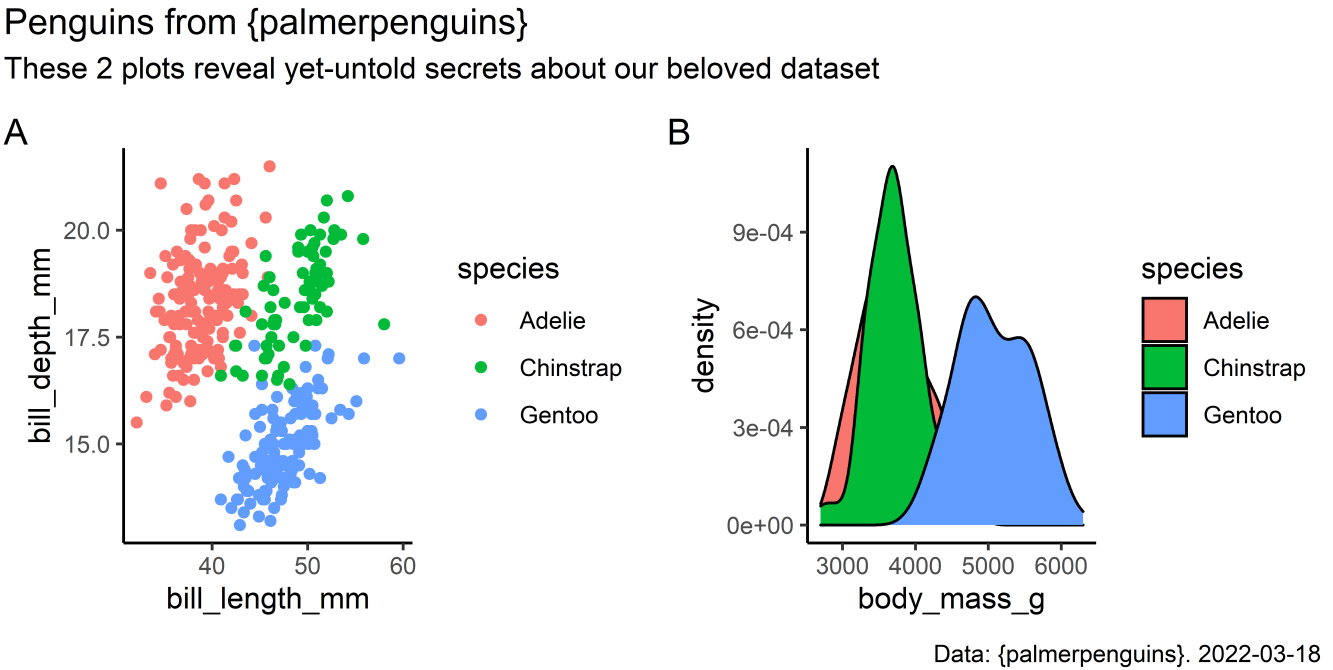
```
library(patchwork)

p1 <- ggplot(penguins) + geom_point(aes(bill_length_mm, bill_depth_mm, color = species))
p2 <- ggplot(penguins) + geom_density(aes(body_mass_g, fill = species))

p1 + p2 +
  plot_annotation(
    tag_levels = 'A',
    title = "Penguins from {palmerpenguins}",
    subtitle = "These 2 plots reveal yet-untold secrets about our beloved dataset",
    caption = "Data: {palmerpenguins}. 2022-03-18")
```

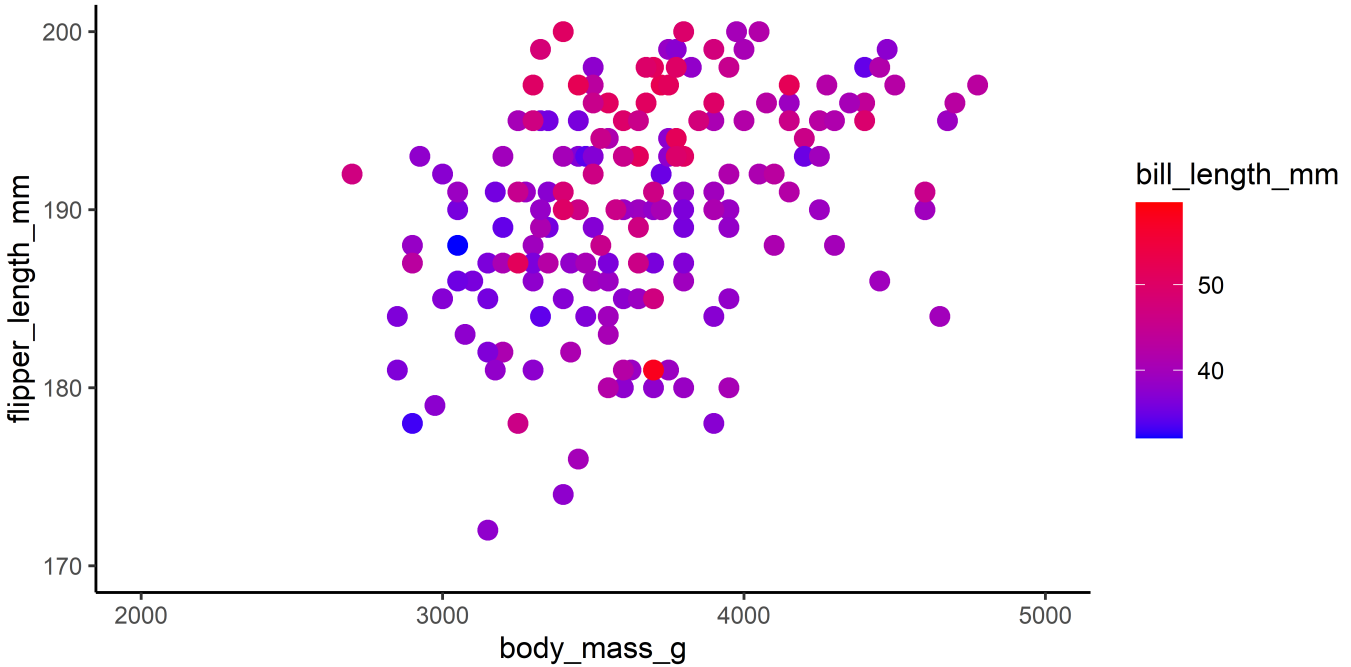
Комбинирование графиков

```
p1 + p2 +
  plot_annotation(
    tag_levels = 'A',
    title = "Penguins from {palmerpenguins}",
    subtitle = "These 2 plots reveal yet-untold secrets about our beloved dataset",
    caption = "Data: {palmerpenguins}. 2022-03-18")
```



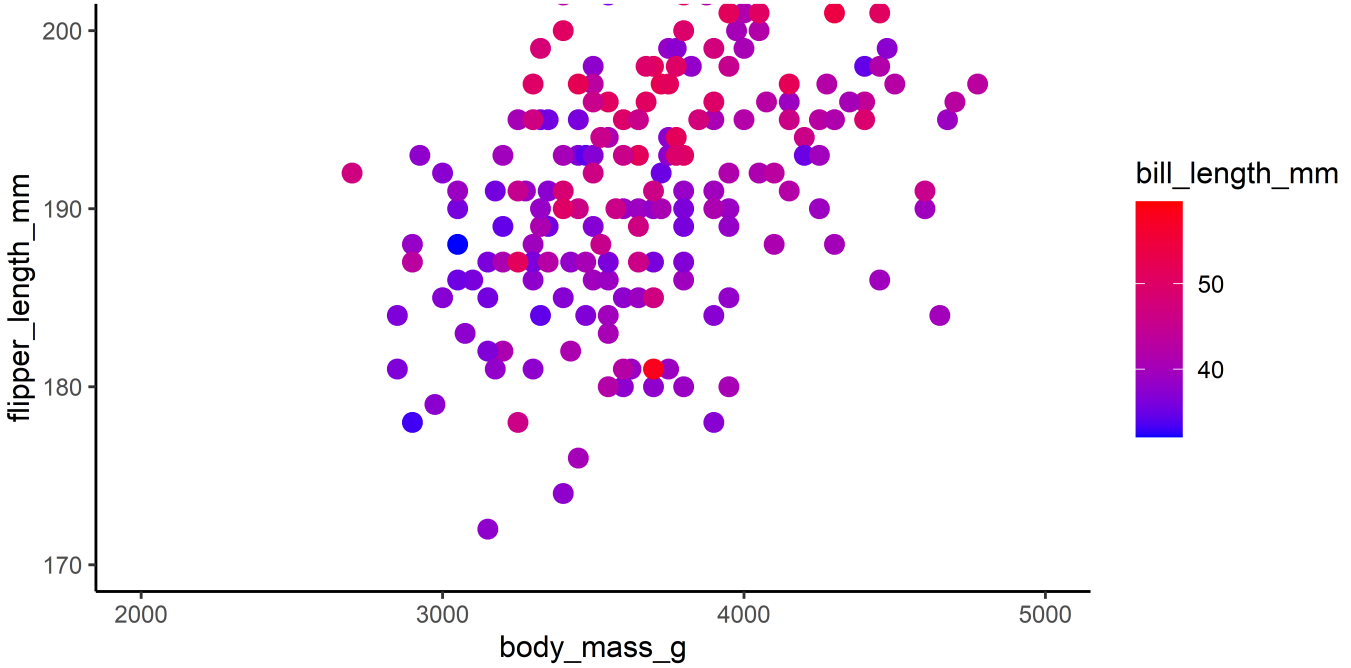
Zoom: xlim & ylim

```
ggplot(penguins, aes(x = body_mass_g, y = flipper_length_mm)) +
  geom_point(aes(color = bill_length_mm), size = 3) +
  scale_color_gradient(low = "blue", high = "red") +
  xlim(2000, 5000) + ylim(170, 200)
```



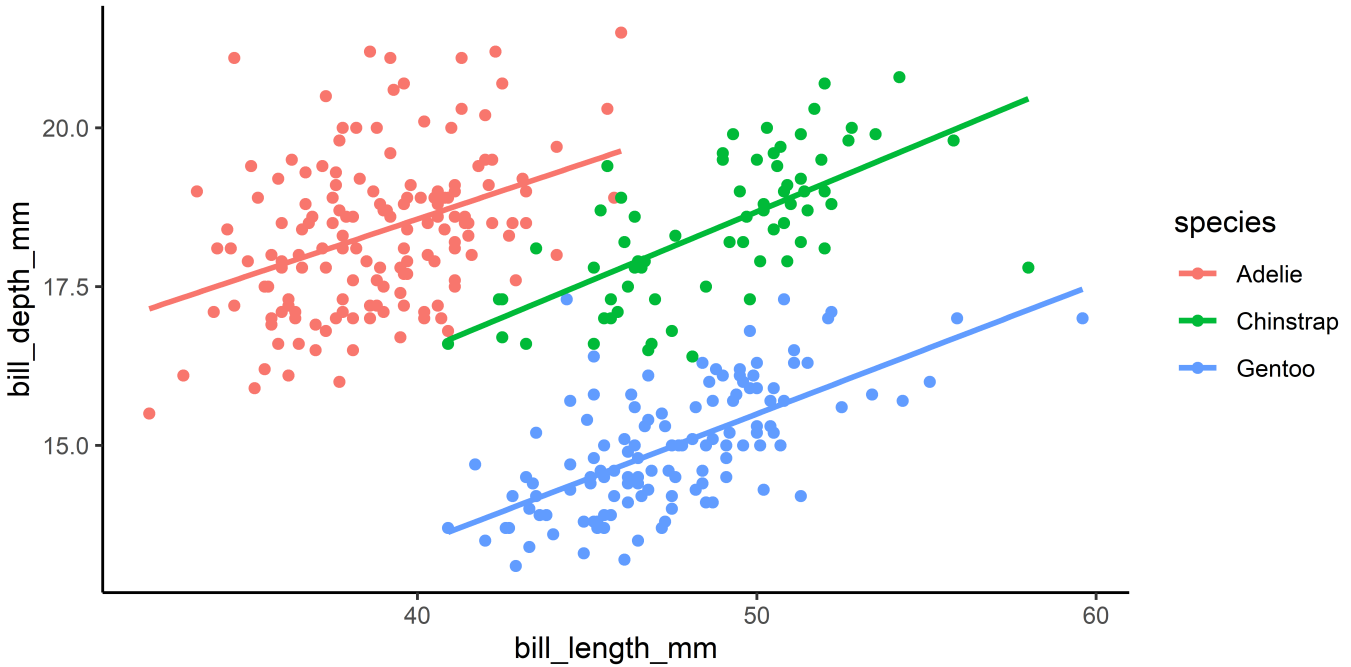
Безопасный zoom

```
ggplot(penguins, aes(x = body_mass_g, y = flipper_length_mm)) +
  geom_point(aes(color = bill_length_mm), size = 3) +
  scale_color_gradient(low = "blue", high = "red") +
  coord_cartesian(xlim = c(2000, 5000), ylim = c(170, 200))
```



Линия тренда

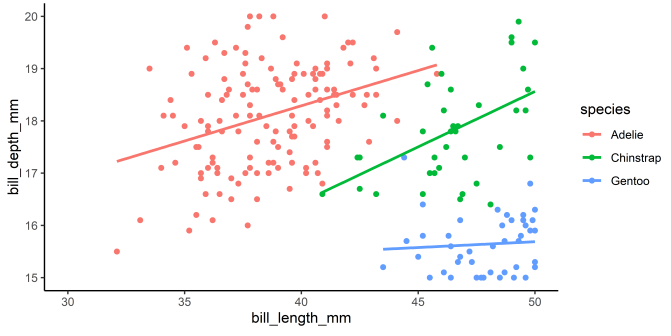
```
ggplot(penguins, aes(bill_length_mm, bill_depth_mm, color = species)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE)
```



Zoom

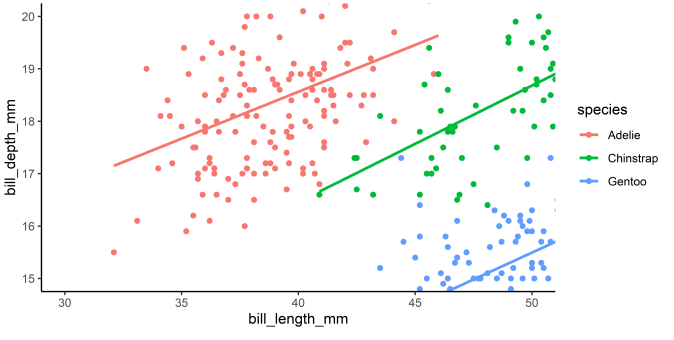
Плохо

```
ggplot(penguins, aes(bill_length_mm,
    bill_depth_mm, color = species)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  xlim(30, 50) + ylim(15, 20)
```



Хорошо

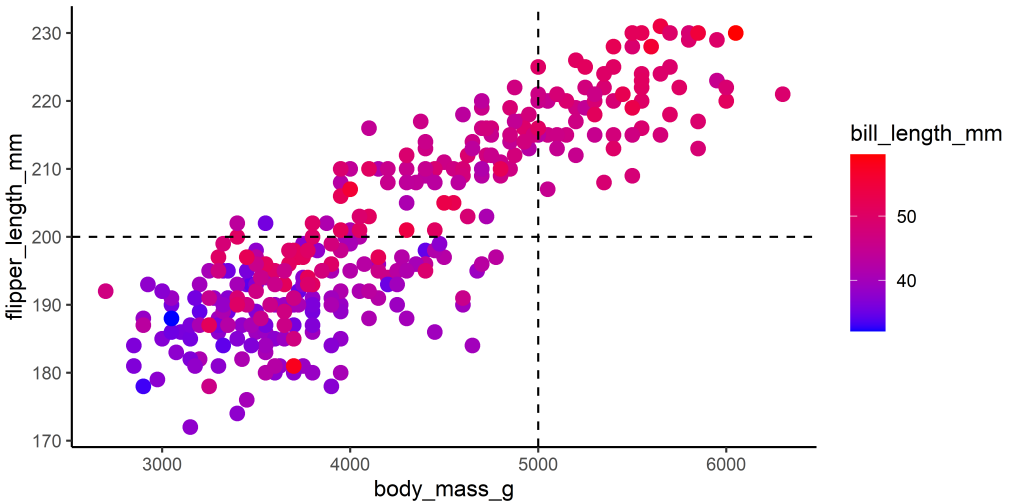
```
ggplot(penguins, aes(bill_length_mm,
    bill_depth_mm, color = species)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  coord_cartesian(xlim = c(30, 50), ylim = c(15,
    20))
```



Дополнительные линии

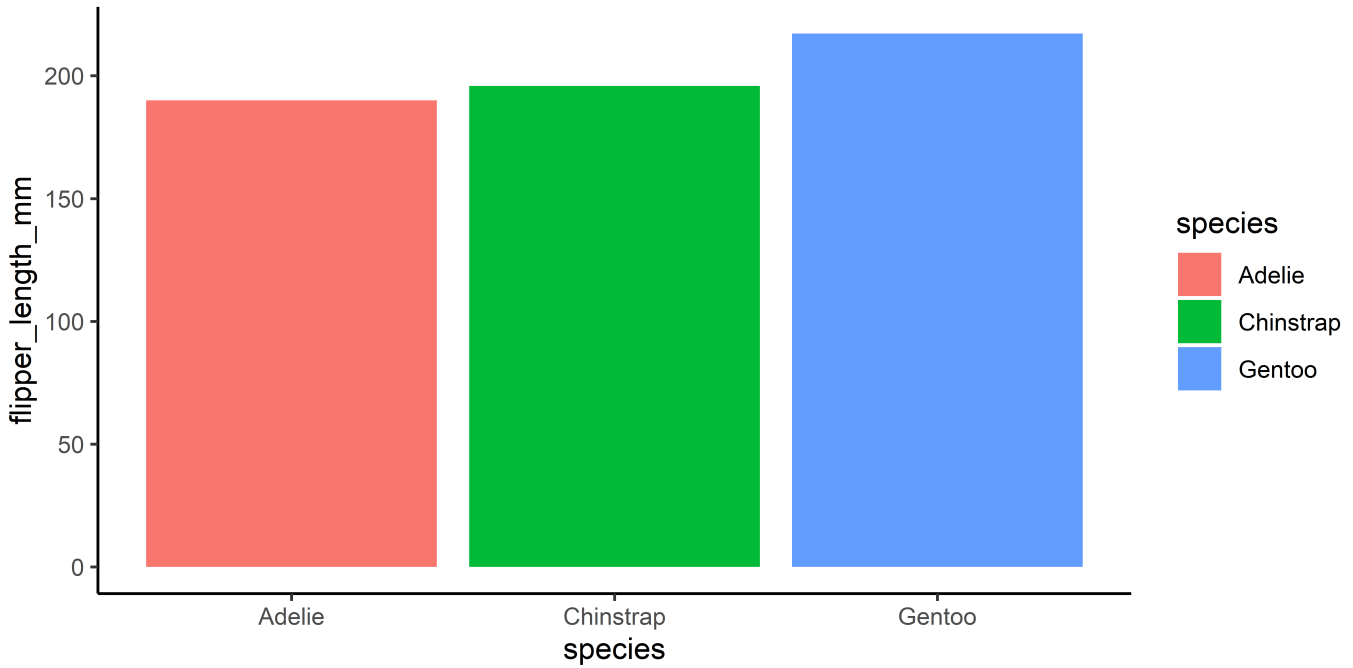
- `geom abline`
- `geom hline`
- `geom_vline`

```
ggplot(penguins, aes(x = body_mass_g, y = flipper_length_mm)) +
  geom_point(aes(color = bill_length_mm), size = 3) +
  scale_color_gradient(low = "blue", high = "red") +
  geom_vline(xintercept = 5000, linetype = "dashed") +
  geom_hline(yintercept = 200, linetype = "dashed")
```



Barplot - summary statistic

```
ggplot(penguins) +
  geom_bar(aes(x = species, y = flipper_length_mm, fill = species),
           stat = "summary", fun = "mean")
```



Barplot & errorbar

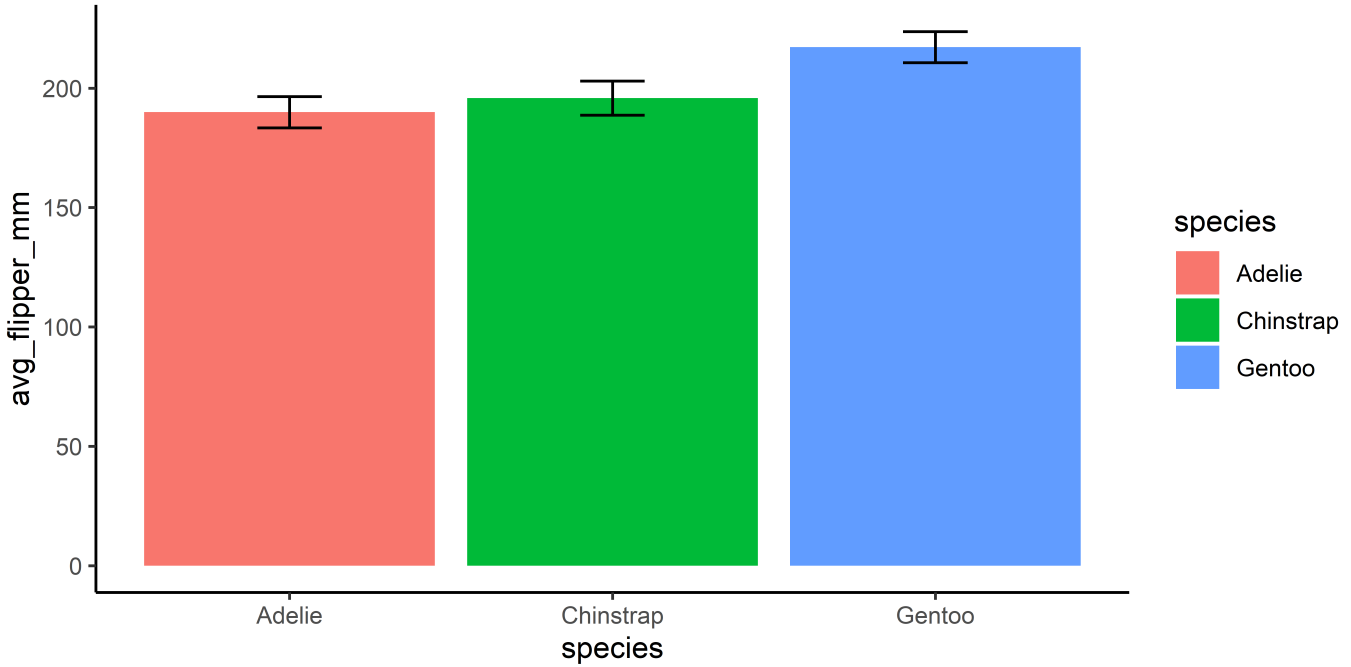
```
penguins_stat <- penguins %>%
  group_by(species) %>%
  summarise(
    avg_flipper_mm = mean(flipper_length_mm, na.rm = TRUE),
    min_flipper_mm = avg_flipper_mm - sd(flipper_length_mm, na.rm = TRUE),
    max_flipper_mm = avg_flipper_mm + sd(flipper_length_mm, na.rm = TRUE))

penguins_stat
```

```
# A tibble: 3 x 4
  species   avg_flipper_mm min_flipper_mm max_flipper_mm
  <chr>              <dbl>          <dbl>          <dbl>
1 Adelie              190.           183.           196.
2 Chinstrap           196.           189.           203.
3 Gentoo              217.           211.           224.
```

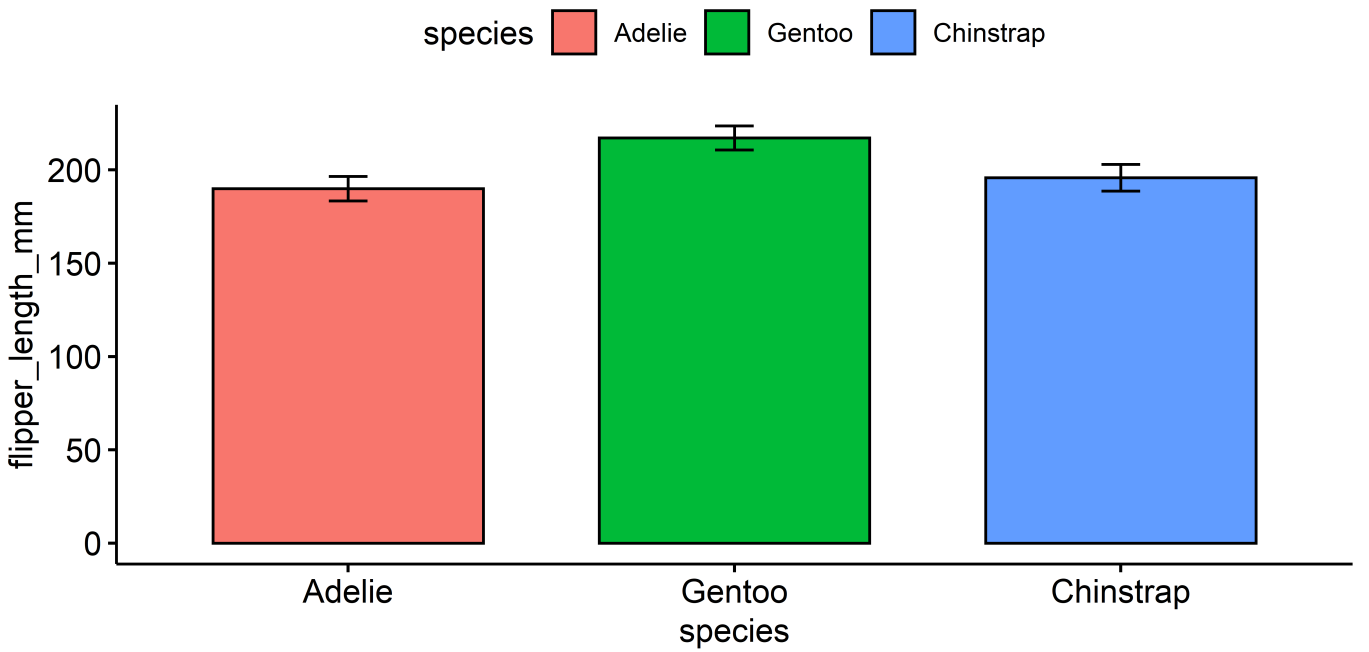
Barplot & errorbar

```
ggplot(penguins_stat) +
  geom_bar(aes(x = species, y = avg_flipper_mm, fill = species),
           stat = "identity") +
  geom_errorbar(aes(x = species, ymin = min_flipper_mm, ymax = max_flipper_mm), width = 0.2)
```



Альтернатива - пакет ggpubr

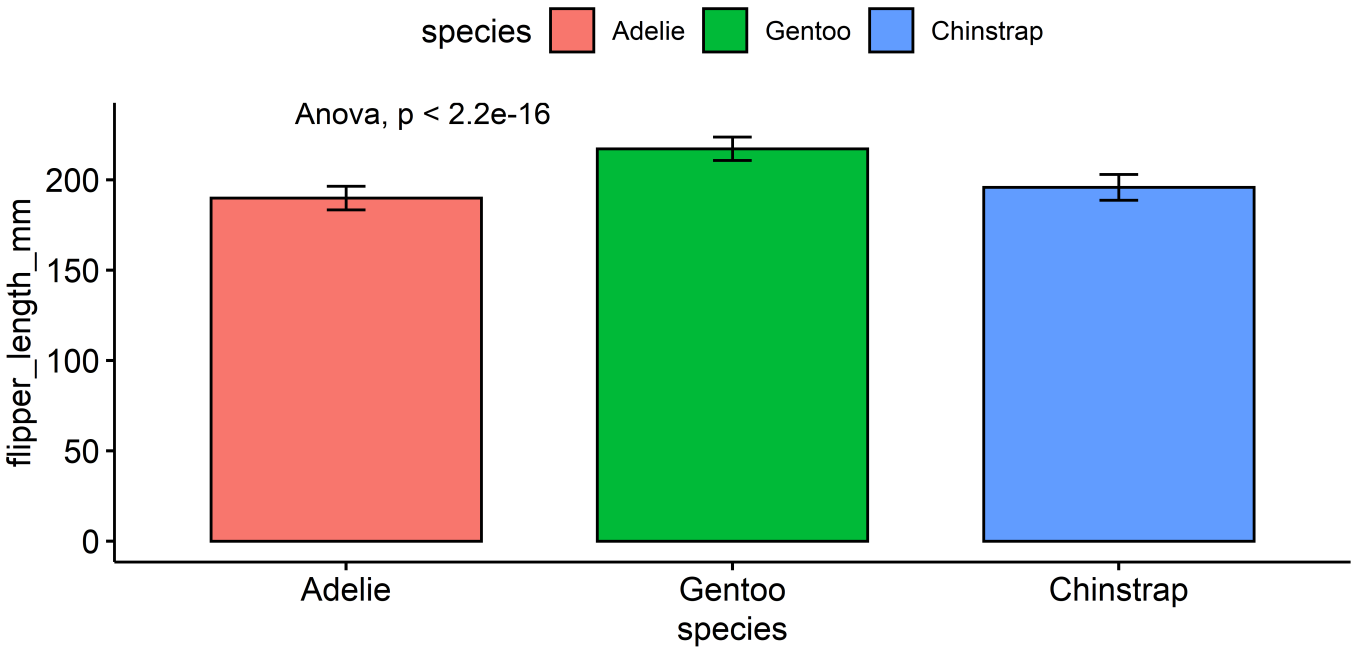
```
library(ggpubr)
ggbarplot(penguins, x = "species", y = "flipper_length_mm", fill = "species", add = "mean_sd")
```



ggpubr

Результаты статистических тестов

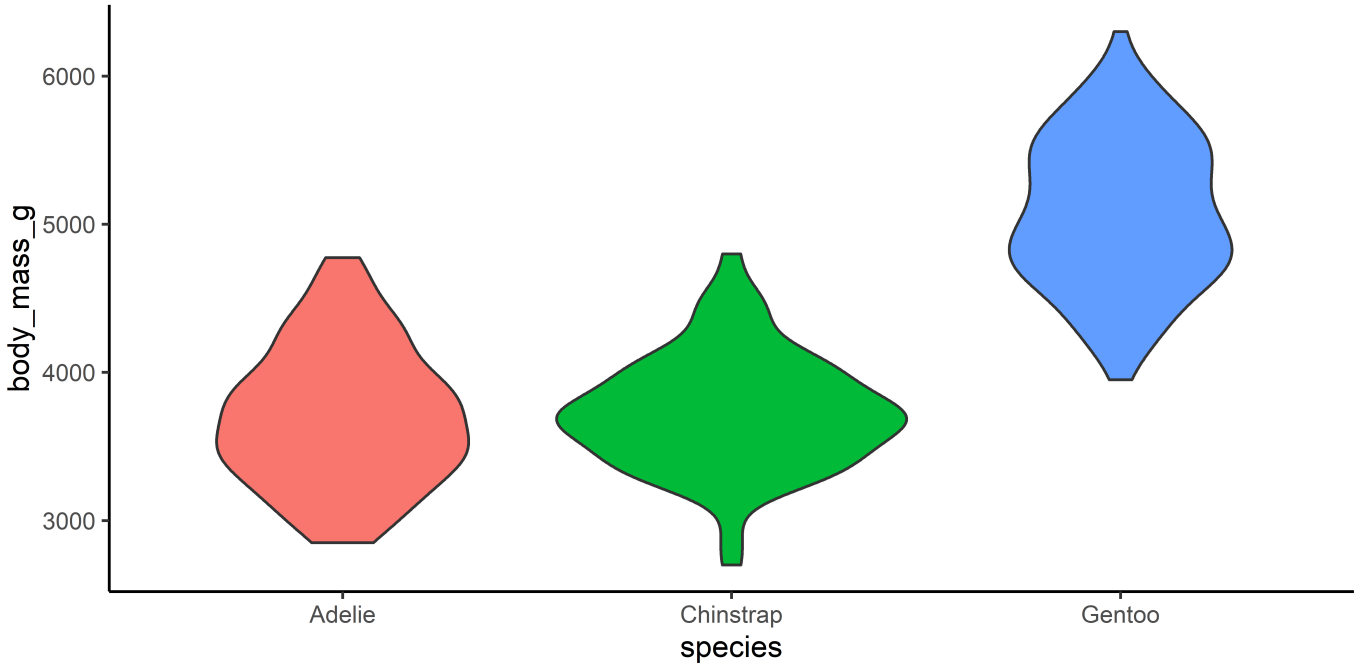
```
ggbarplot(penguins, x = "species", y = "flipper_length_mm", fill = "species", add = "mean_sd") +
  stat_compare_means(method = "anova")
```



Скрипичная диаграмма

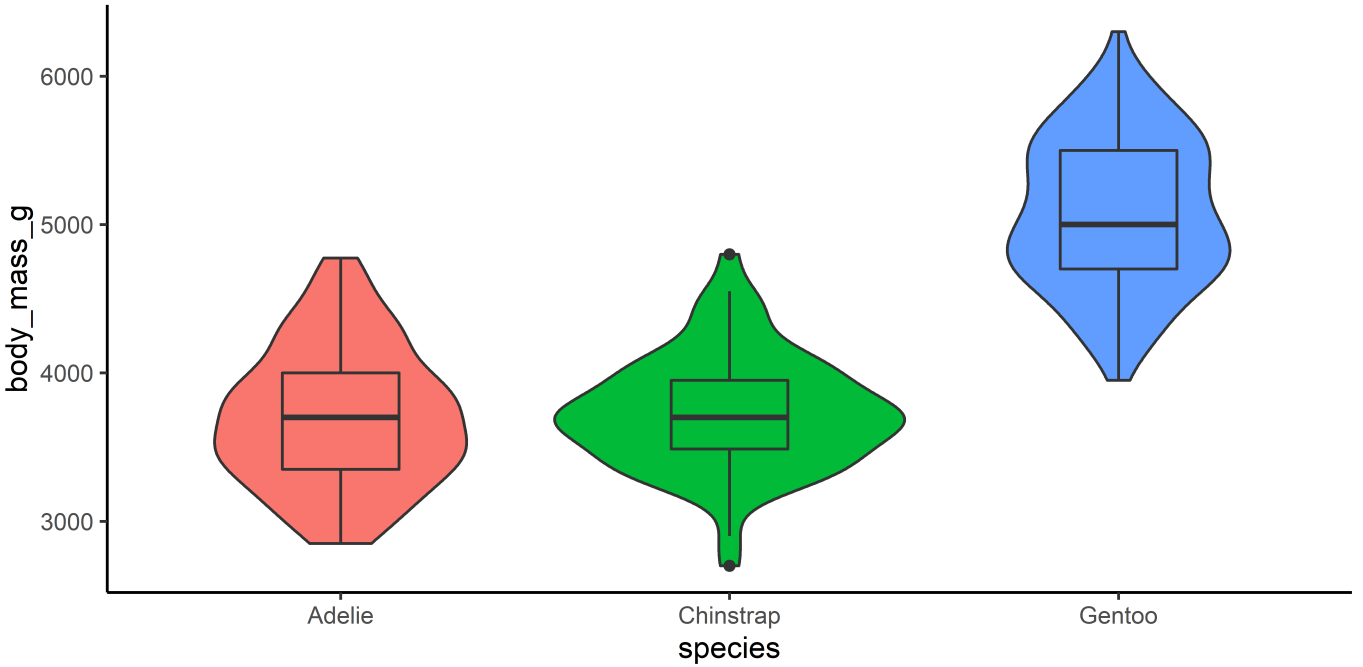
Violin plot

```
ggplot(penguins) +
  geom_violin(aes(x = species, y = body_mass_g, fill = species)) +
  theme(legend.position = "none")
```



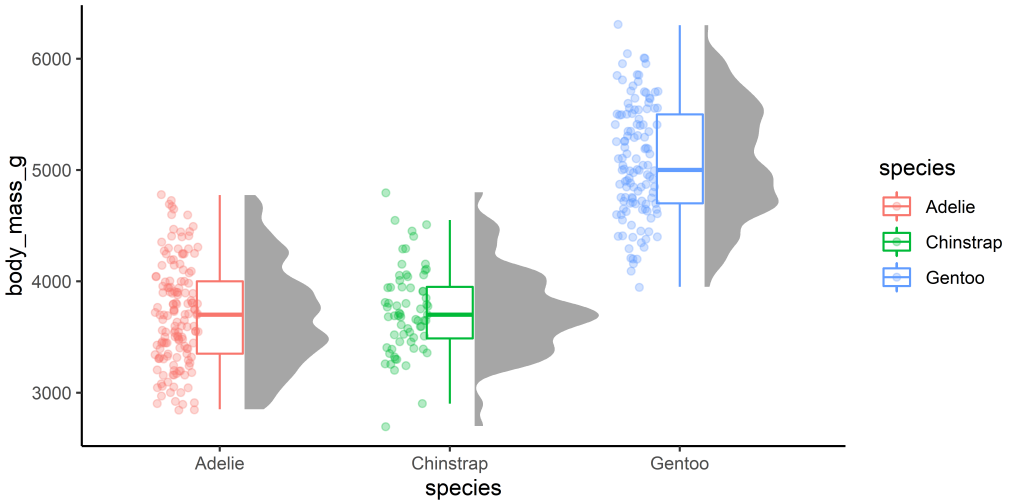
Комбинирование geoms

```
ggplot(penguins, aes(x = species, y = body_mass_g, fill = species)) +
  geom_violin() +
  geom_boxplot(width = 0.3) +
  theme(legend.position = "none")
```



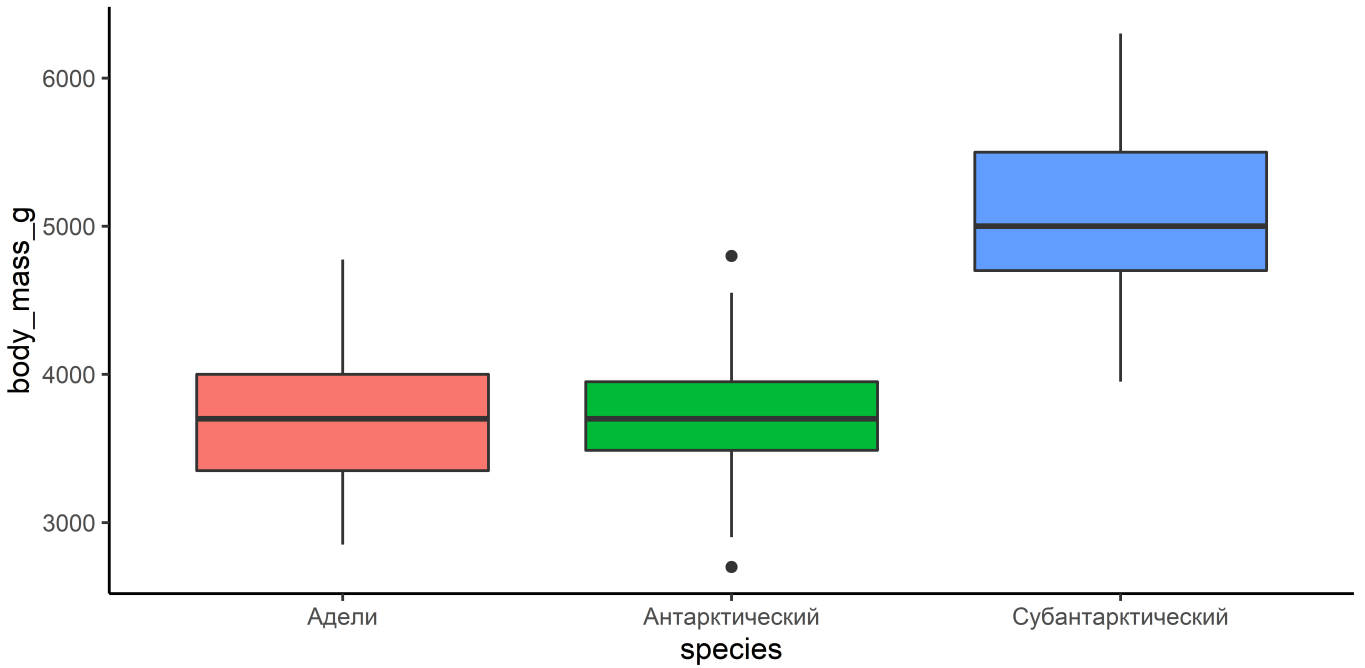
Raincloud plot

```
ggplot(penguins, aes(x = species, y = body_mass_g, color = species)) +
  # половинка violin
  ggdist::stat_halfeye(
    adjust = 0.5, width = 0.6, .width = 0,
    justification = -0.2, point_colour = NA) +
  # боксплот
  geom_boxplot(width = 0.2, outlier.shape = NA) +
  # точки
  gghalves::geom_half_point(side = "l", range_scale = 0.5, alpha = 0.3)
```



Breaks & labels

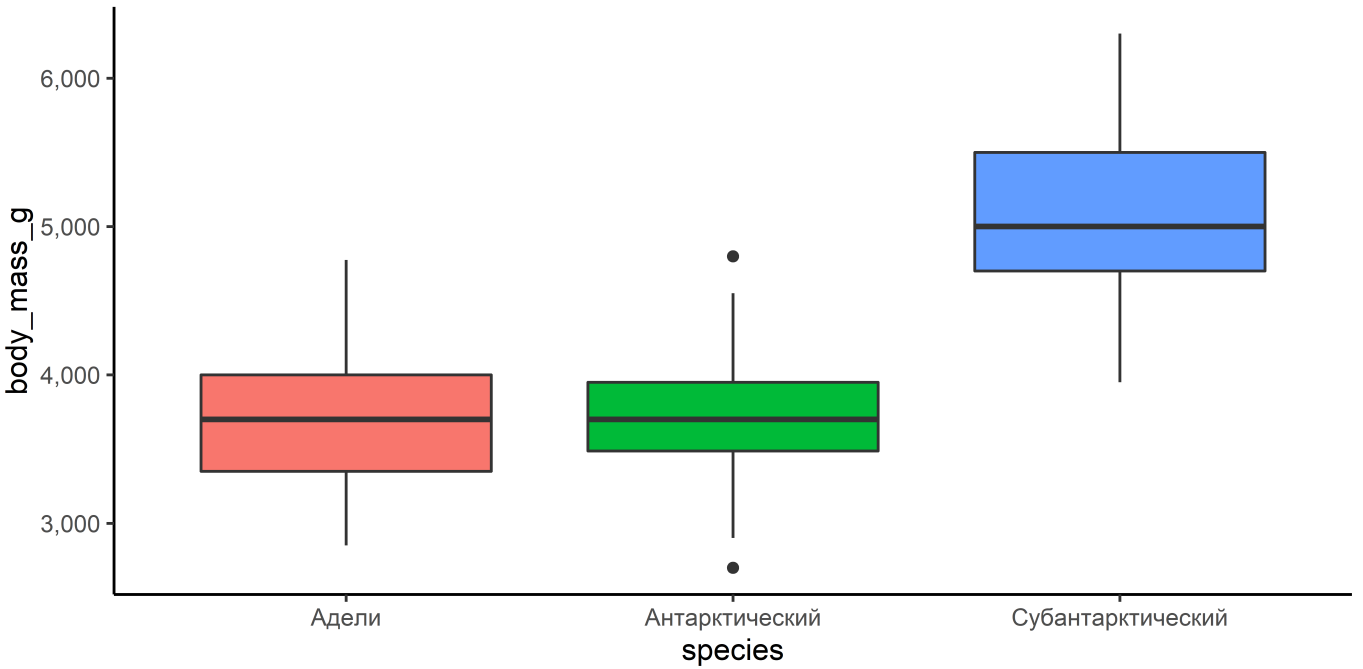
```
ggplot(penguins, aes(x = species, y = body_mass_g, fill = species)) +
  geom_boxplot() +
  theme(legend.position = "none") +
  scale_x_discrete(breaks = c("Adelie", "Chinstrap", "Gentoo"),
                   labels =  c("Адели", "Антарктический", "Субантарктический"))
```



Формат значений по осям

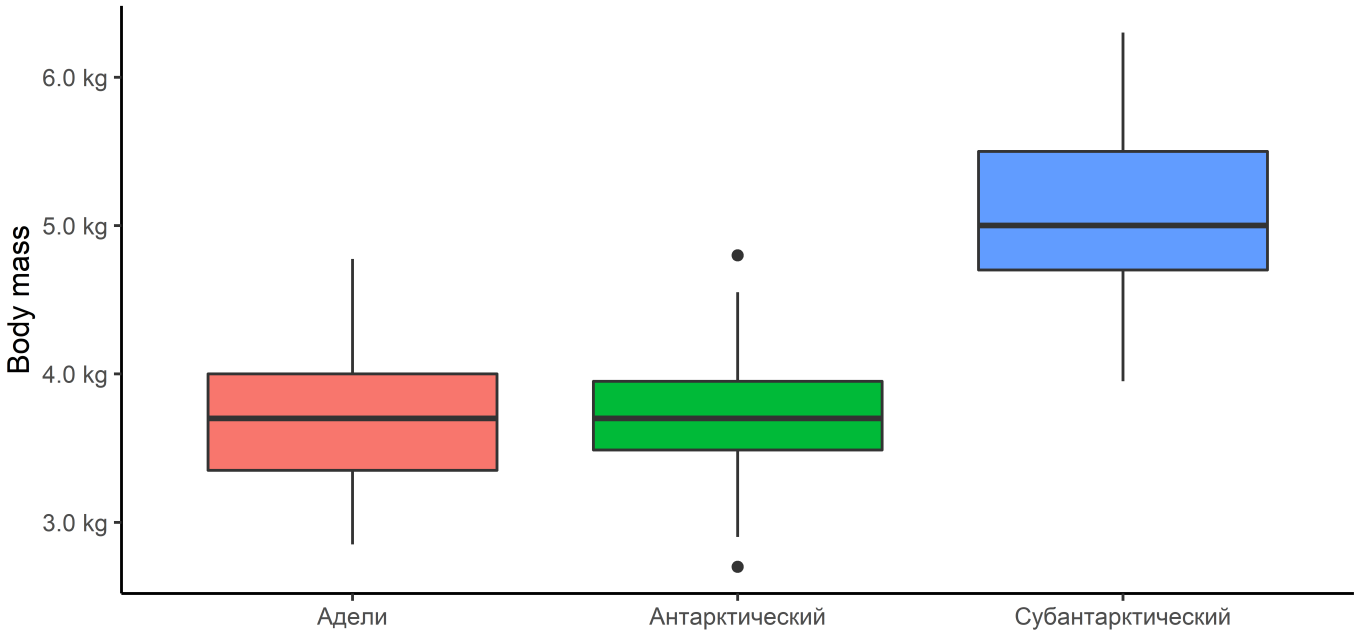
[scales::label_number()]

```
ggplot(penguins, aes(x = species, y = body_mass_g, fill = species)) +
  geom_boxplot() +
  theme(legend.position = "none") +
  scale_x_discrete(breaks = c("Adelie", "Chinstrap", "Gentoo"),
                   labels =  c("Адели", "Антарктический", "Субантарктический")) +
  scale_y_continuous(labels = scales::comma_format())
```



Формат значений по осям

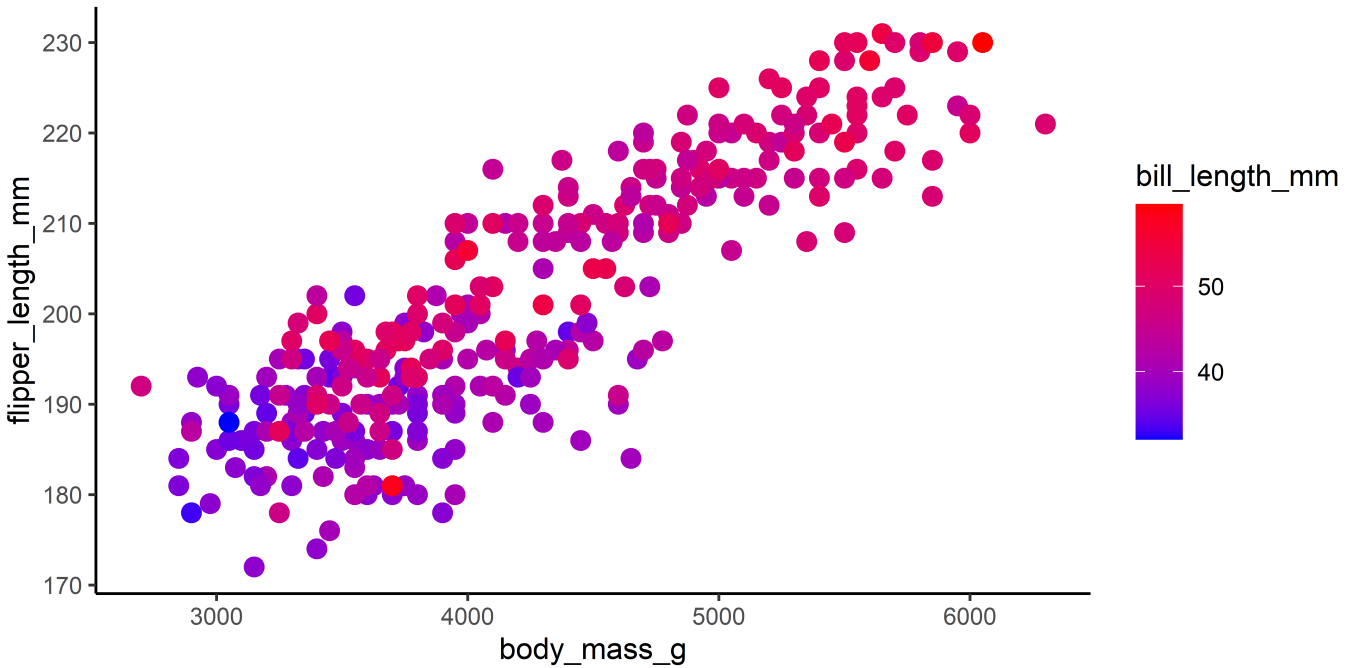
```
ggplot(penguins, aes(x = species, y = body_mass_g, fill = species)) +
  geom_boxplot() +
  theme(legend.position = "none") +
  scale_x_discrete(breaks = c("Adelie", "Chinstrap", "Gentoo"),
                   labels =  c("Адели", "Антарктический", "Субантарктический")) +
  labs(x = "", y = "Body mass") +
  scale_y_continuous(labels = scales::comma_format(scale = 1/1000, suffix = " kg"))
```



Градиенты

- `scale color gradient`
- `scale color gradient2`
- `scale_color_gradientn`

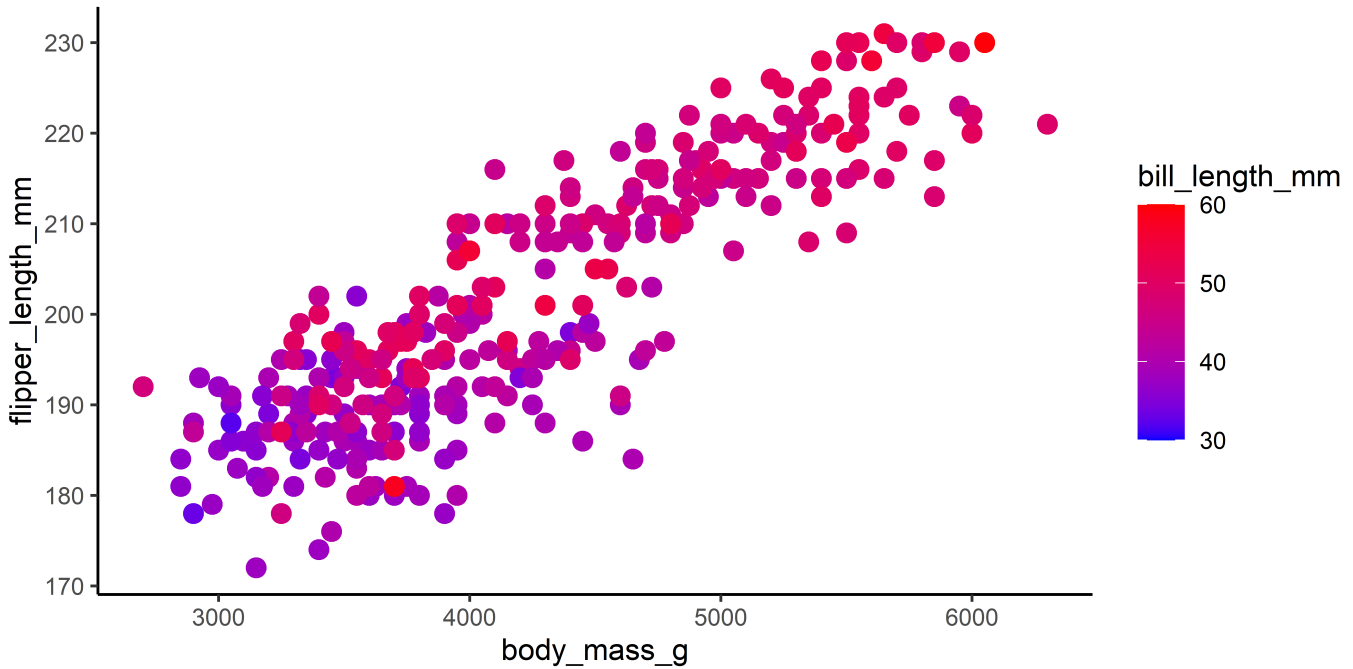
```
ggplot(penguins, aes(x = body_mass_g, y = flipper_length_mm)) +
  geom_point(aes(color = bill_length_mm), size = 3) +
  scale_color_gradient(low = "blue", high = "red")
```



Градиенты: breaks & labels

- `scale color gradient`
- `scale color gradient2`
- `scale_color_gradientn`

```
ggplot(penguins, aes(x = body_mass_g, y = flipper_length_mm)) +
  geom_point(aes(color = bill_length_mm), size = 3) +
  scale_color_gradient(low = "blue", high = "red",
    breaks = c(30, 40, 50, 60), limits = c(30, 60))
```



Цветовые палитры

R пакеты

- {ggplot2} - viridis
- {RColorBrewer}
- {tvthemes}
- {ggsci}
- {MetBrewer}
- {wesanderson}
- ...

Сайты

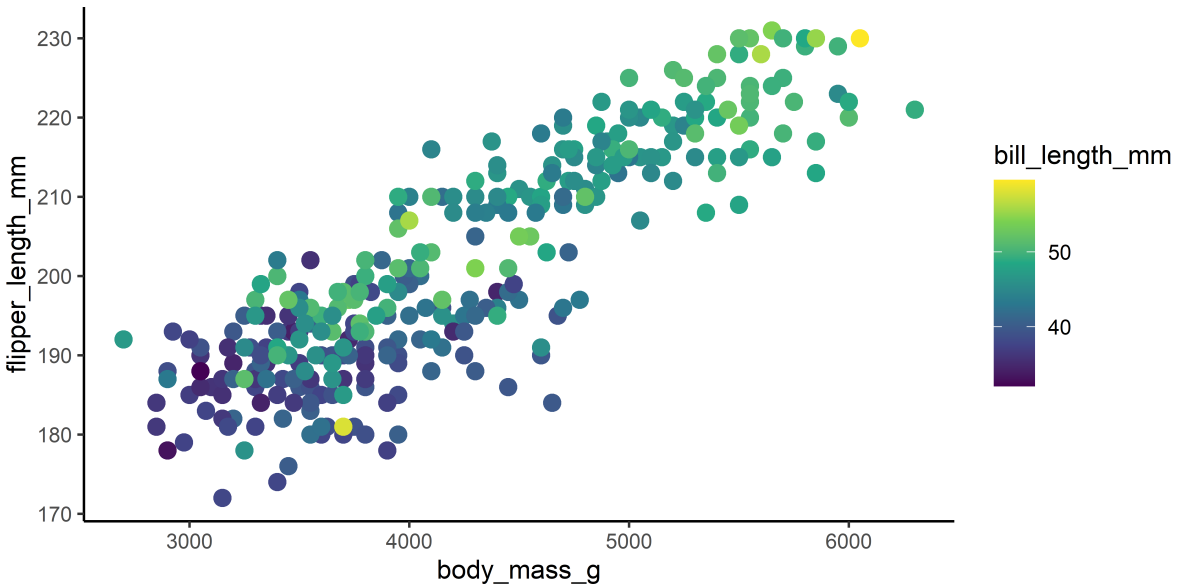
- coolors.co
- ...

Цветовые палитры

viridis

- `scale color viridis c` для непрерывных переменных
- `scale color viridis d` для дискретных переменных
- `scale_color_viridis_b` для бинированных переменных

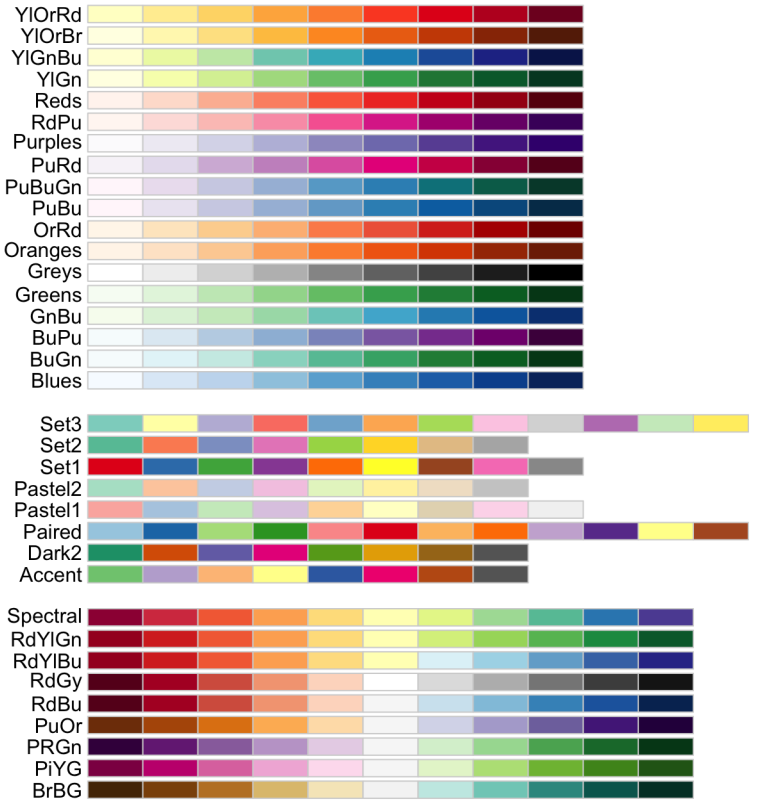
```
ggplot(penguins, aes(x = body_mass_g, y = flipper_length_mm)) +
  geom_point(aes(color = bill_length_mm), size = 3) +
  scale_color_viridis_c()
```



Цветовые палитры

RColorBrewer

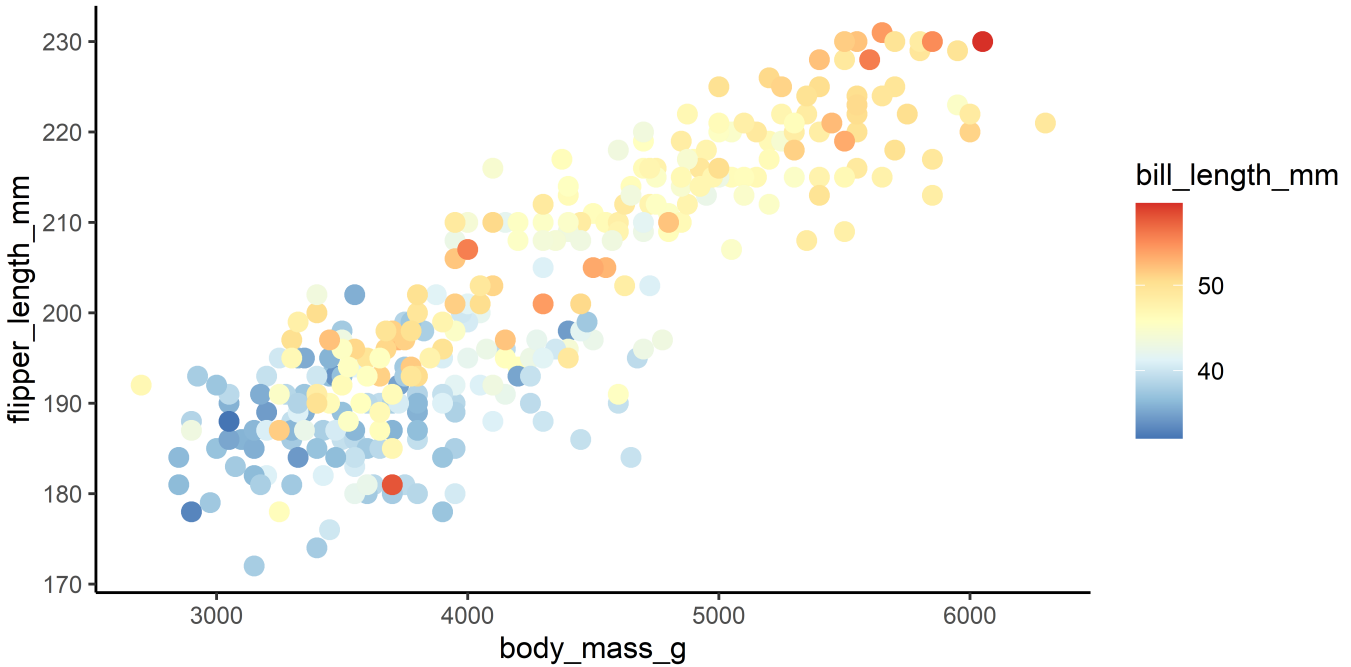
- `scale color brewer` для дискретных переменных
- `scale_color_distiller` для непрерывных переменных



Цветовые палитры

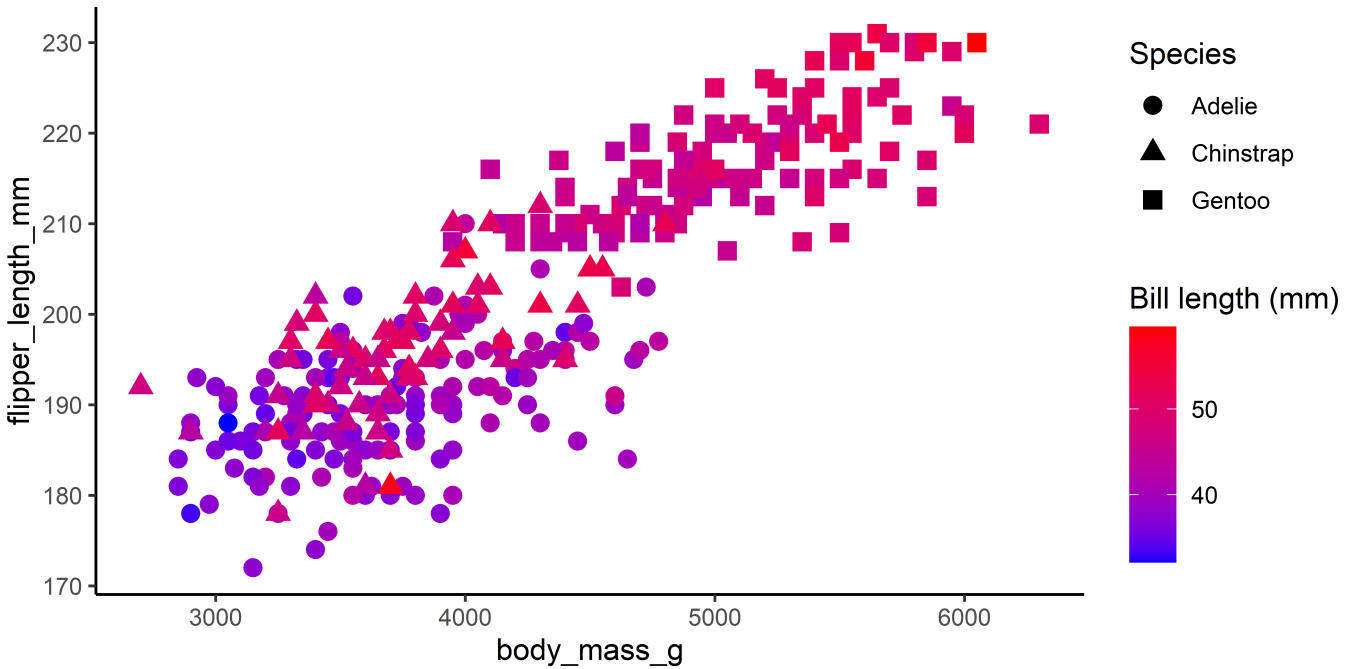
RColorBrewer

```
ggplot(penguins, aes(x = body_mass_g, y = flipper_length_mm)) +
  geom_point(aes(color = bill_length_mm), size = 3) +
  scale_color_distiller(palette = "RdYlBu", direction = -1)
```



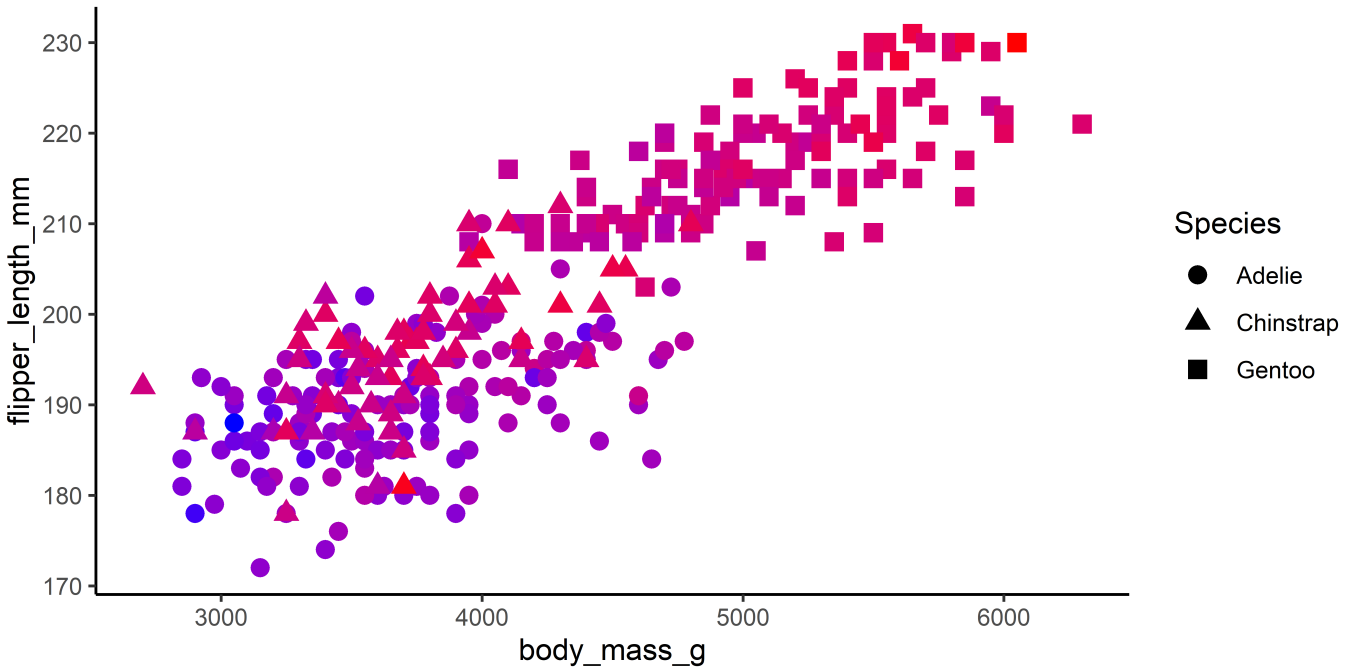
Несколько легенд

```
ggplot(penguins, aes(x = body_mass_g, y = flipper_length_mm)) +
  geom_point(aes(color = bill_length_mm, shape = species), size = 3) +
  guides(
    shape = guide_legend(order = 1, title = "Species"),
    color = guide_colorbar(order = 2, title = "Bill length (mm)")) +
  scale_color_gradient(low = "blue", high = "red")
```



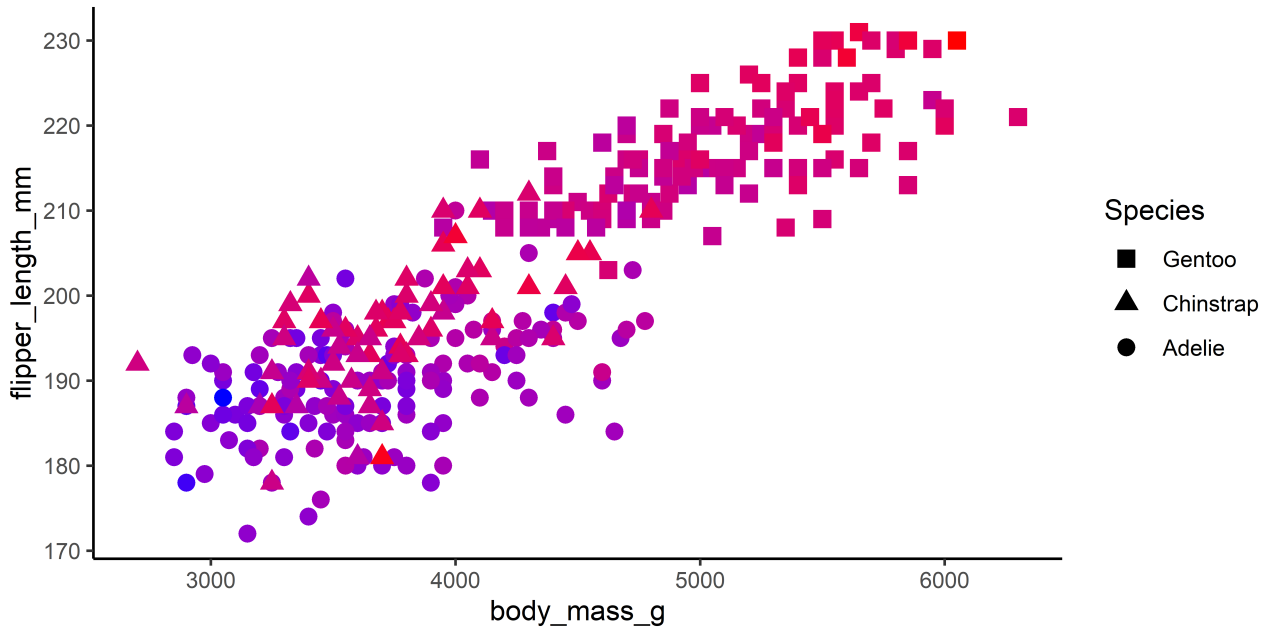
Несколько легенд

```
ggplot(penguins, aes(x = body_mass_g, y = flipper_length_mm)) +
  geom_point(aes(color = bill_length_mm, shape = species), size = 3) +
  guides(
    shape = guide_legend(order = 1, title = "Species"),
    color = "none") +
  scale_color_gradient(low = "blue", high = "red")
```



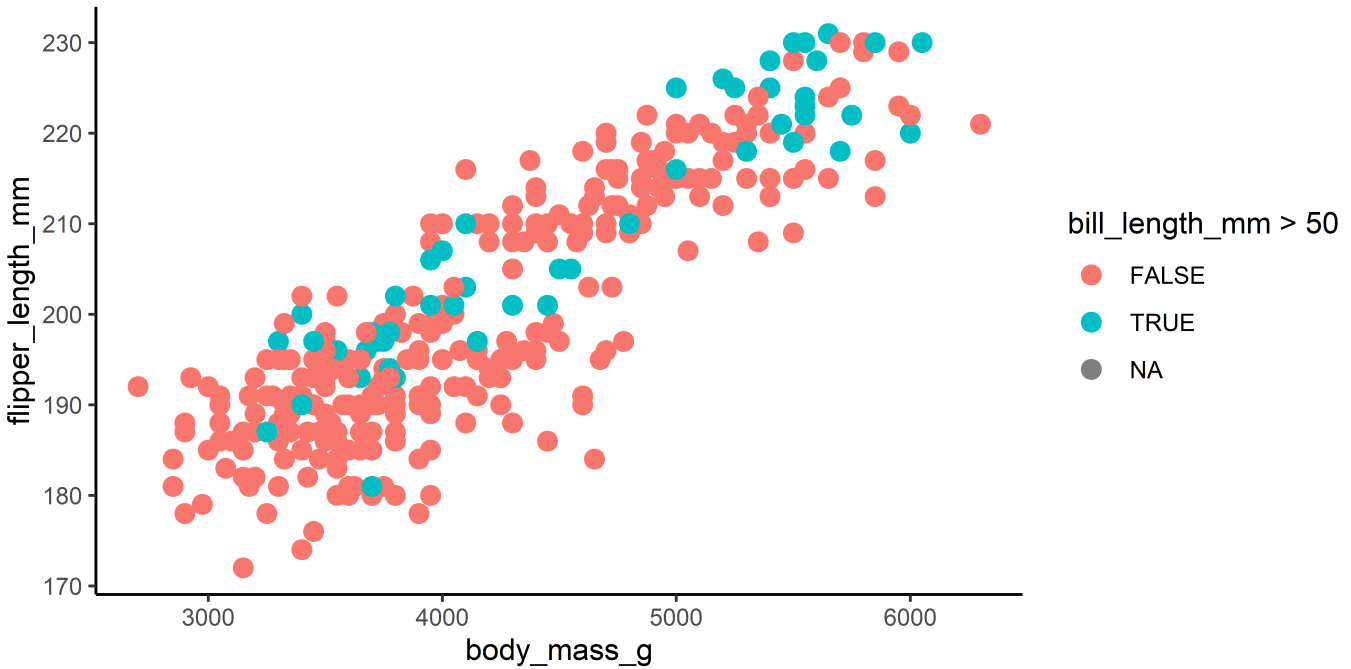
Несколько легенд

```
ggplot(penguins, aes(x = body_mass_g, y = flipper_length_mm)) +
  geom_point(aes(color = bill_length_mm, shape = species), size = 3) +
  guides(
    shape = guide_legend(order = 1, title = "Species", reverse = TRUE),
    color = "none") +
  scale_color_gradient(low = "blue", high = "red")
```



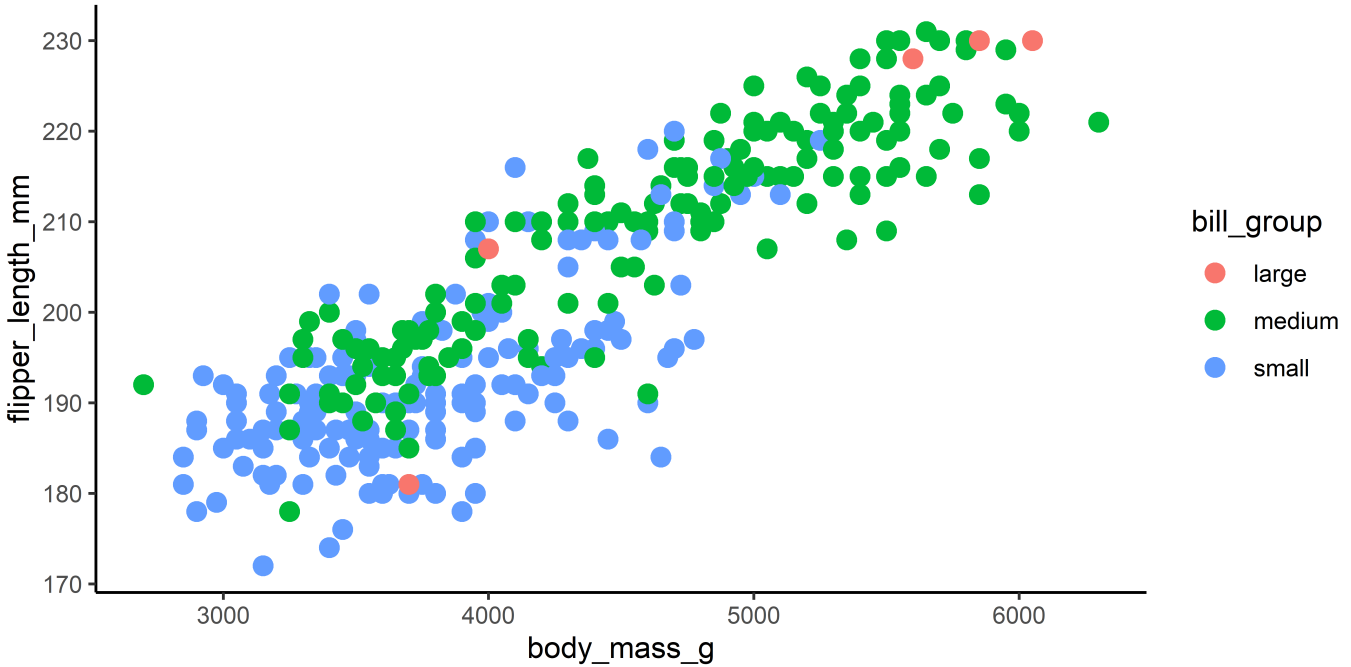
Цвет по условию

```
ggplot(penguins, aes(x = body_mass_g, y = flipper_length_mm)) +
  geom_point(aes(color = bill_length_mm > 50), size = 3)
```



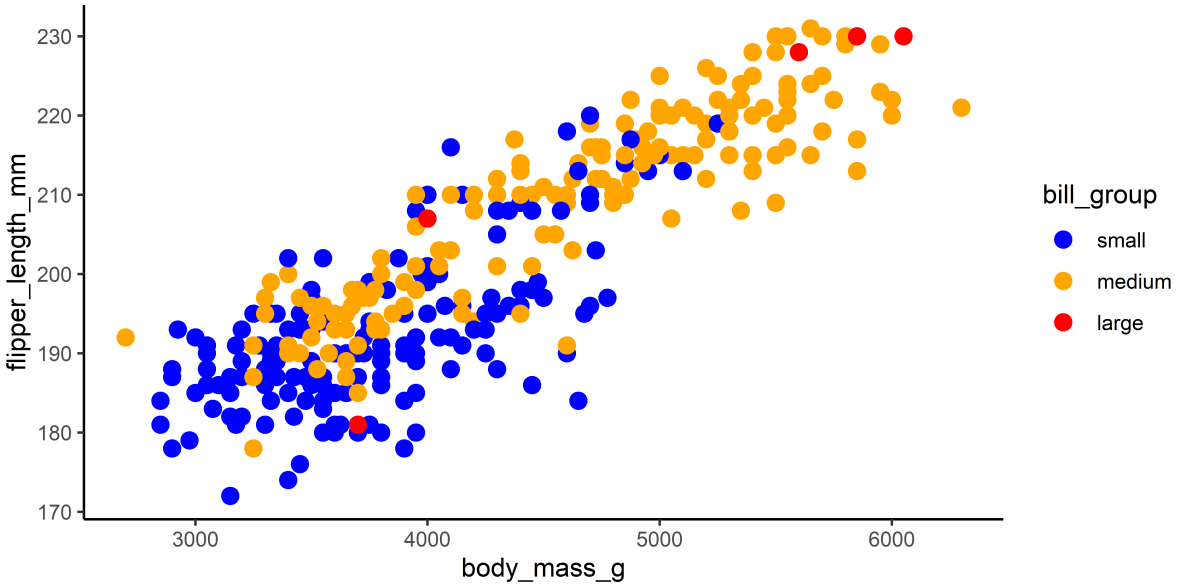
Цвет по условию...

```
penguins %>%
  mutate(bill_group = case_when(
    (bill_length_mm < 45) ~ "small",
    (bill_length_mm >= 55) ~ "large",
    TRUE ~ "medium")) %>%
  ggplot(aes(x = body_mass_g, y = flipper_length_mm)) +
  geom_point(aes(color = bill_group), size = 3)
```



Цвет по условию...

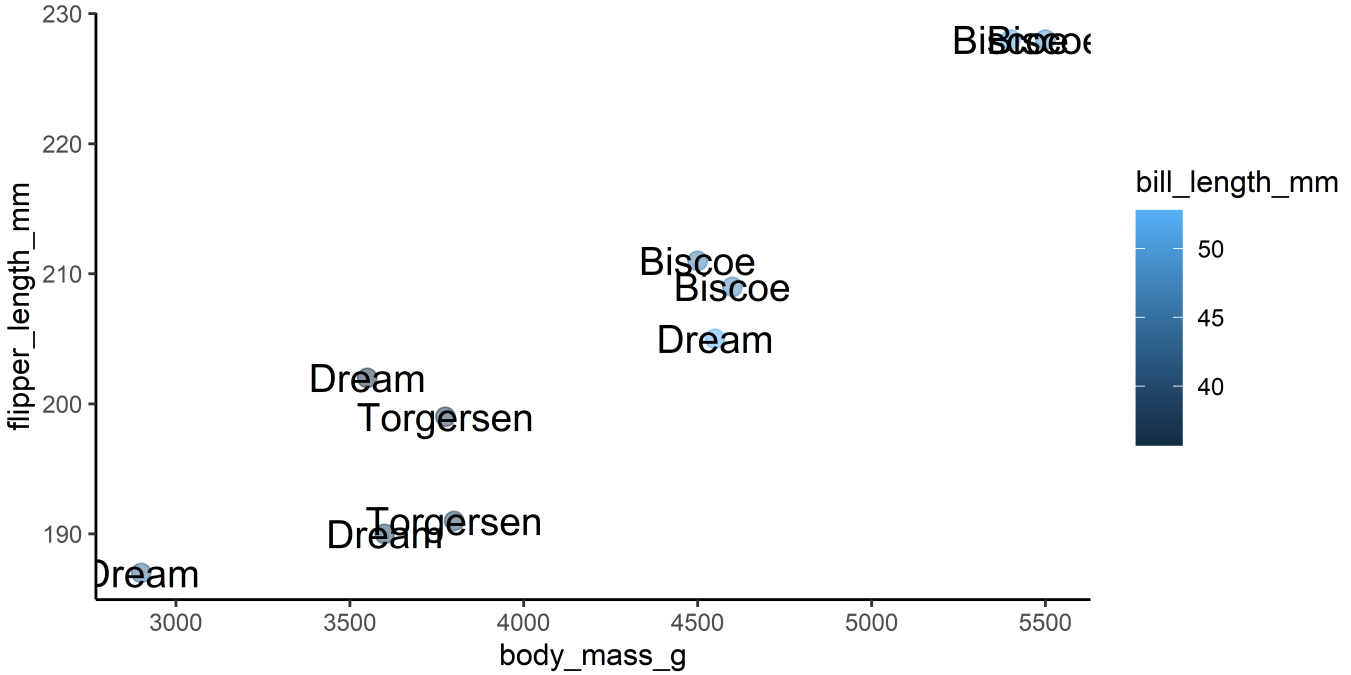
```
penguins %>%
  mutate(bill_group = case_when(
    (bill_length_mm < 45) ~ "small",
    (bill_length_mm >= 55) ~ "large",
    TRUE ~ "medium")) %>%
  ggplot(aes(x = body_mass_g, y = flipper_length_mm)) +
  geom_point(aes(color = bill_group), size = 3) +
  scale_color_manual(values = c("small" = "blue", "medium" = "orange", "large" = "red"))
```



Подписи

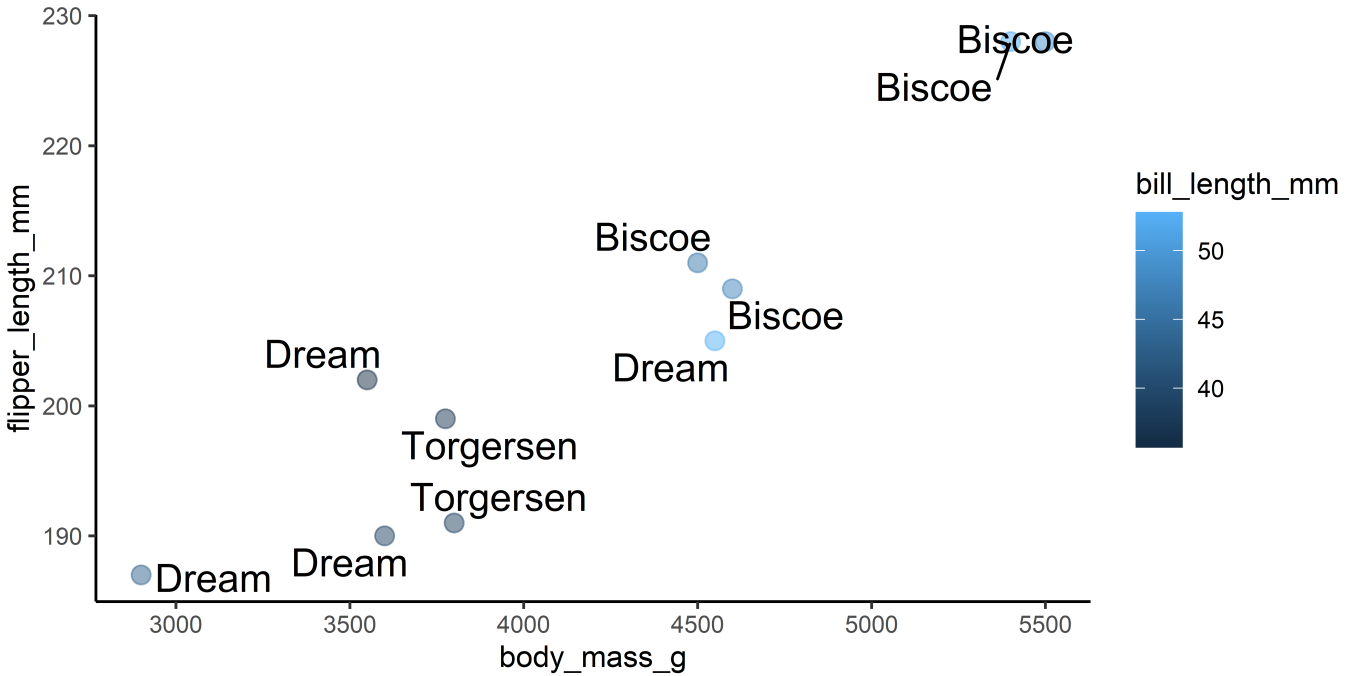
```r
sample_10 <- slice_sample(penguins, n = 10)

ggplot(sample_10, aes(x = body_mass_g, y = flipper_length_mm)) +
  geom_point(aes(color = bill_length_mm), alpha = 0.5, size = 3) +
  # Наследует aes X и Y
  geom_text(mapping = aes(label = island), size = 5)
```



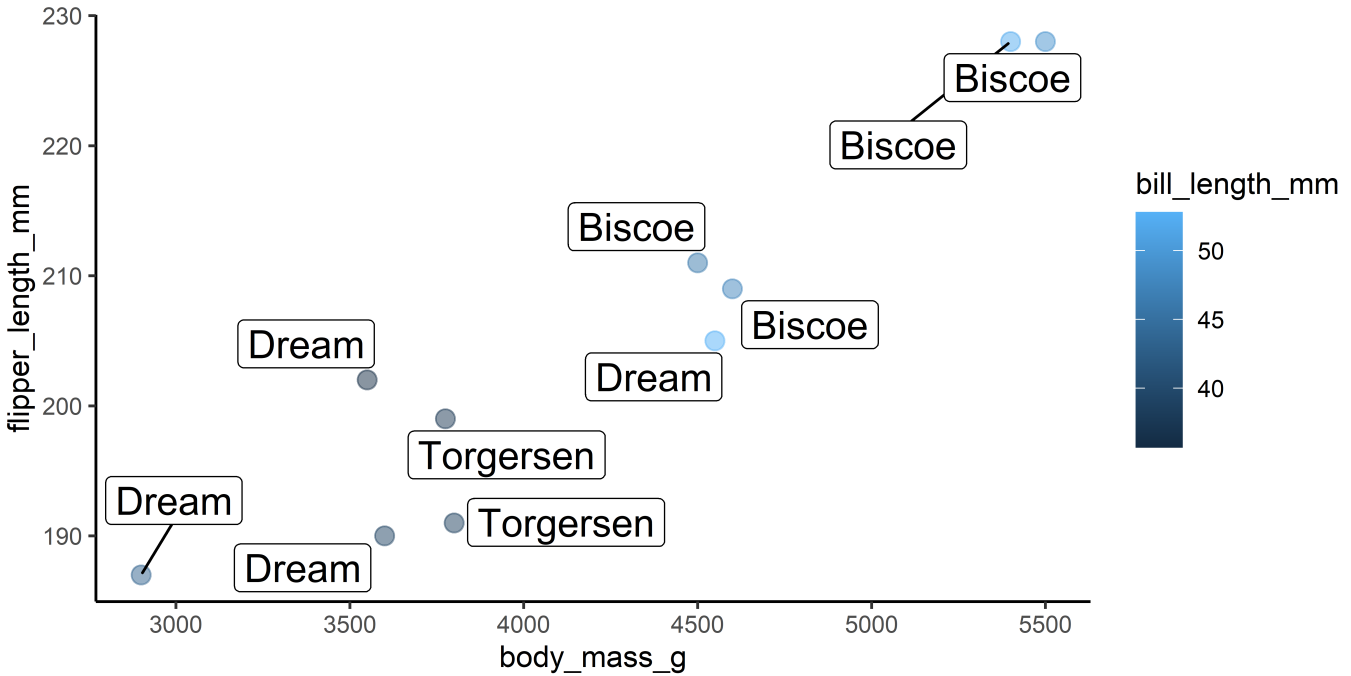
Подписи с {ggrepel}

```
ggplot(sample_10, aes(x = body_mass_g, y = flipper_length_mm)) +
  geom_point(aes(color = bill_length_mm), alpha = 0.5, size = 3) +
  ggrepel::geom_text_repel(mapping = aes(label = island), size = 5)
```



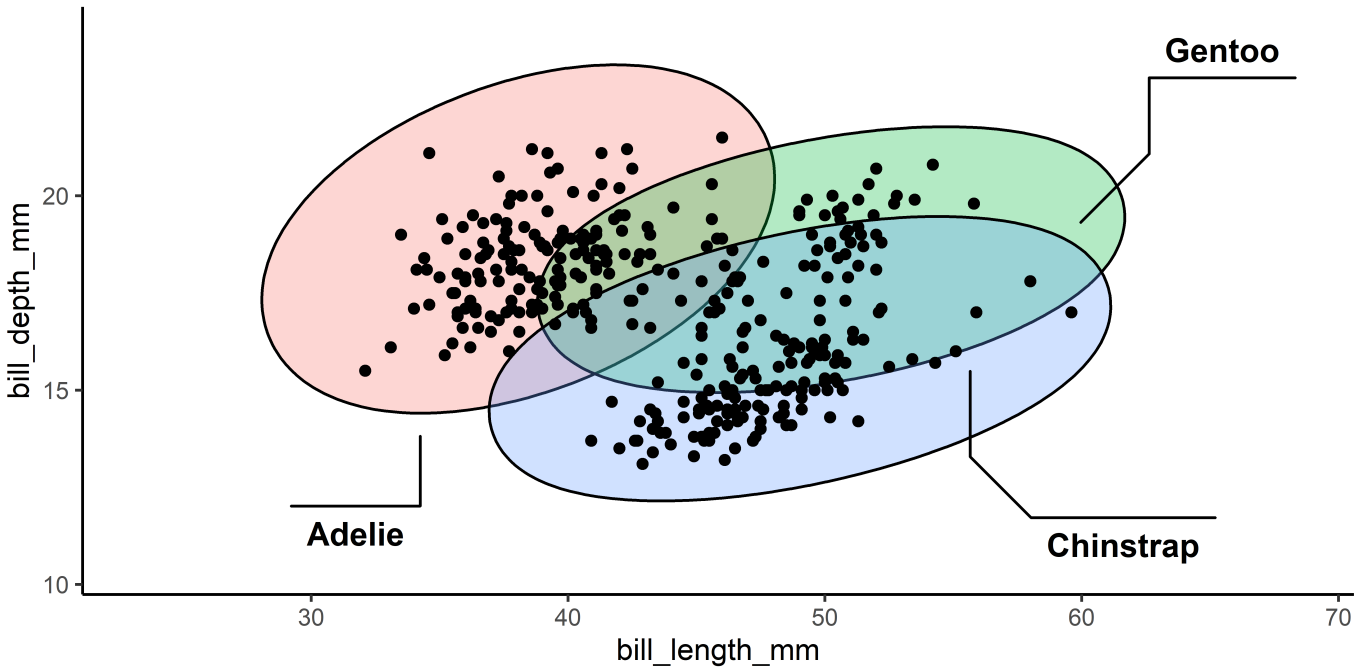
Подписи с {ggrepel}

```
ggplot(sample_10, aes(x = body_mass_g, y = flipper_length_mm)) +
  geom_point(aes(color = bill_length_mm), alpha = 0.5, size = 3) +
  ggrepel::geom_label_repel(mapping = aes(label = island), size = 5)
```



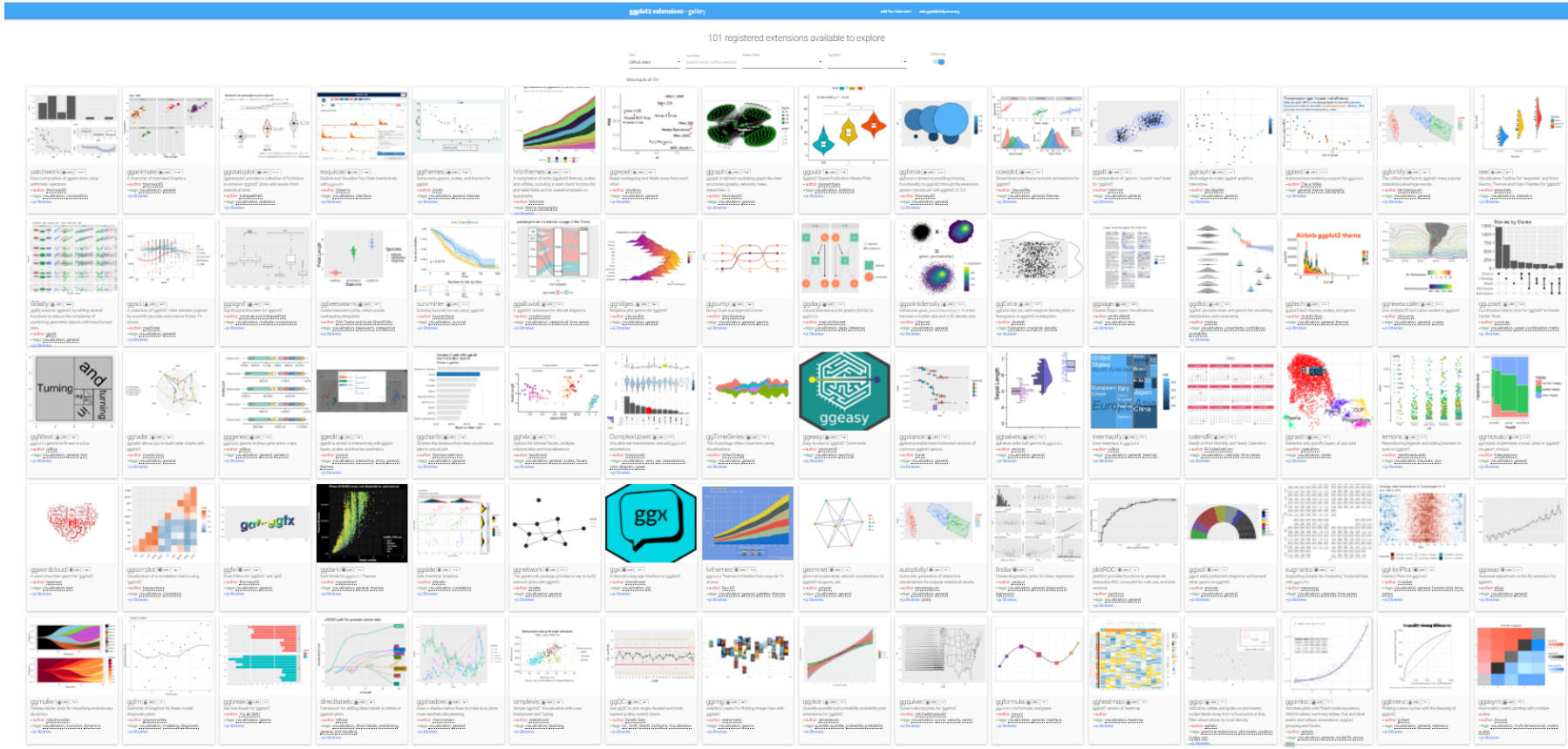
Аннотация с {ggforce}

```
drop_na(penguins, bill_length_mm, bill_depth_mm) %>%
  ggplot(aes(bill_length_mm, bill_depth_mm)) +
  ggforce::geom_mark_ellipse(aes(fill = species, label = species)) +
  geom_point() + theme(legend.position = "none") +
  scale_x_continuous(expand = expansion(mult = .4)) +
  scale_y_continuous(expand = expansion(mult = .4))
```



Ggplot2 extensions

https://exts.ggplot2.tidyverse.org/



Шрифты

Пакеты: {showtext}, {extrafont}, ...

Чтобы сработало в RMarkdown: `{r fig.showtext=TRUE}`

```
library(showtext)
## Загрузить шрифты из Google fonts (https://fonts.google.com/)
font_add_google("Gochi Hand", "gochi")
font_add_google("Schoolbell", "bell")

## Использовать showtext автоматически
showtext_auto()

ggplot(penguins, aes(bill_length_mm, bill_depth_mm, color = species)) +
  geom_point() +
  labs(x = "Bill length (mm)", y = "Bill depth (mm)",
       title = "Penguins from Antarctics") +
  theme(
    text =  element_text(family = "bell", size = 14),
    plot.title = element_text(family = "gochi", size = rel(1.6), hjust = 0.5))
```