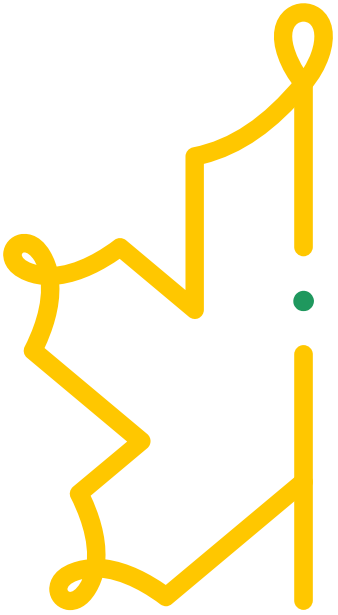
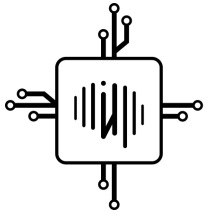


Razвед
анализ
данных

Tidyverse

как смотреть на данные



Фонд
интеллект

Анастасия Жарикова

Лекция 8 - 2022

Набор данных про тыквы

Задача

Есть набор данных

Нужно каким-то образом изучить эти данные

Что делать?

Задача

Есть набор данных

Нужно каким-то образом изучить эти данные

Что делать?

Обзор

Сколько наблюдений? Сколько переменных? Какого типа переменные?

```
pump <- read_delim('pump.tab')
glimpse(pump)
```

```
Rows: 25,833
Columns: 13
$ id                <chr> "2013-S", "2013-S", "2013-S", "2013-S", "2013-S", "2…
$ place             <chr> "1", "2", "3", "4", "5", "6", "7", "8", "9", "10", "…
$ weight_lbs        <dbl> 1264.0, 1233.0, 1175.0, 1159.5, 1094.0, 1093.0, 1064…
$ grower_name       <chr> "Pierpont, Edwin", "Hunt, Jane & Phil", "Holub, Scot…
$ city              <chr> "Jefferson", "Cameron", "Eugene", "Shawville", "Ston…
$ state_prov        <chr> "Maine", "Ontario", "Oregon", "Quebec", "Ontario", "…
$ country           <chr> "United States", "Canada", "United States", "Canada"…
$ gpc_site          <chr> "Damariscotta Pumpkinfest and Regatta Weigh-off", "P…
$ seed_mother       <chr> "996 Haist", "996 Haist", "1037.5 Jarvis", "1153 Kli…
$ pollinator_father <chr> "1109 Pierpont", "self", "Self", "800 Neily", "self"…
$ ott               <dbl> 355, 376, 382, 376, 394, 358, 350, 352, 350, 333, 35…
$ est_weight        <dbl> 983, 1154, 1203, 1154, 1304, 1010, 946, 963, 948, 82…
$ pct_chart         <dbl> 29, 7, -2, 0, -16, 8, 13, 7, 8, 12, -6, 9, 3, -8, 4,…
```

variable - description

id - Year-type

place - Place/ranking

weight_lbs - Weight in pounds

grower_name - Name of grower

city - City

state_prov - State/Province

country - Country

gpc_site - GPC site (great pumpkin commonwealth)

seed_mother - Seed mother

pollinator_father - Father

ott - Over the top inches, can be used to estimate weight

est_weight - Estimated weight in lbs

pct_chart - Percent on chart

Подсчет отсутствующих значений

```
apply(pump, 2, function(x) sum(is.na(x)))
```

```
            id          place     weight_lbs    grower_name
             0              0              0              0
          city      state_prov        country       gpc_site
          2565              0              0              0
   seed_mother pollinator_father            ott     est_weight
          7892           9554           3156           3156
     pct_chart
          3156
```

Исследование разнообразия значений в столбце

```
pump %>%
  count(id)
```

```
# A tibble: 51 × 2
   id         n
   <chr>  <int>
 1 2013-S   151
 2 2013-T   289
 3 2013-W   273
 4 2014-F   330
 5 2014-L   185
 6 2014-P  1900
 7 2014-S   192
 8 2014-T   294
 9 2014-W   272
10 2015-F   319
# … with 41 more rows
```

Столбец id

2013-F : год-плод

F - field pumpkin

P - giant pumpkin

S - giant squash

W - giant watermelon

L - long gourd

T - tomato

Разделить столбец на несколько

```
pump %>%
  separate(id, c('year', 'type'), '-') %>% head(2)
```

```
# A tibble: 2 × 14
  year  type  place weight_lbs grower_name     city  state_prov country gpc_site
  <chr> <chr> <chr>      <dbl> <chr>           <chr> <chr>      <chr>   <chr>
1 2013  S     1           1264 Pierpont, Edwin Jeff… Maine      United… Damaris…
2 2013  S     2           1233 Hunt, Jane & P… Came… Ontario    Canada  Port El…
# … with 5 more variables: seed_mother <chr>, pollinator_father <chr>,
#   ott <dbl>, est_weight <dbl>, pct_chart <dbl>
```

```
pump %>%
  separate(id, c('year', 'type'), '-', remove = F) %>% head(2)
```

```
# A tibble: 2 × 15
  id     year  type  place weight_lbs grower_name       city  state_prov country
  <chr>  <chr> <chr> <chr>      <dbl> <chr>             <chr> <chr>      <chr>
1 2013-S 2013  S     1           1264 Pierpont, Edwin   Jeff… Maine      United…
2 2013-S 2013  S     2           1233 Hunt, Jane & Phil Came… Ontario    Canada
# … with 6 more variables: gpc_site <chr>, seed_mother <chr>,
#   pollinator_father <chr>, ott <dbl>, est_weight <dbl>, pct_chart <dbl>
```

```
pump %>% separate(id, c('year', 'type'), '-') -> pumpS
head(pumpS, 2)
```

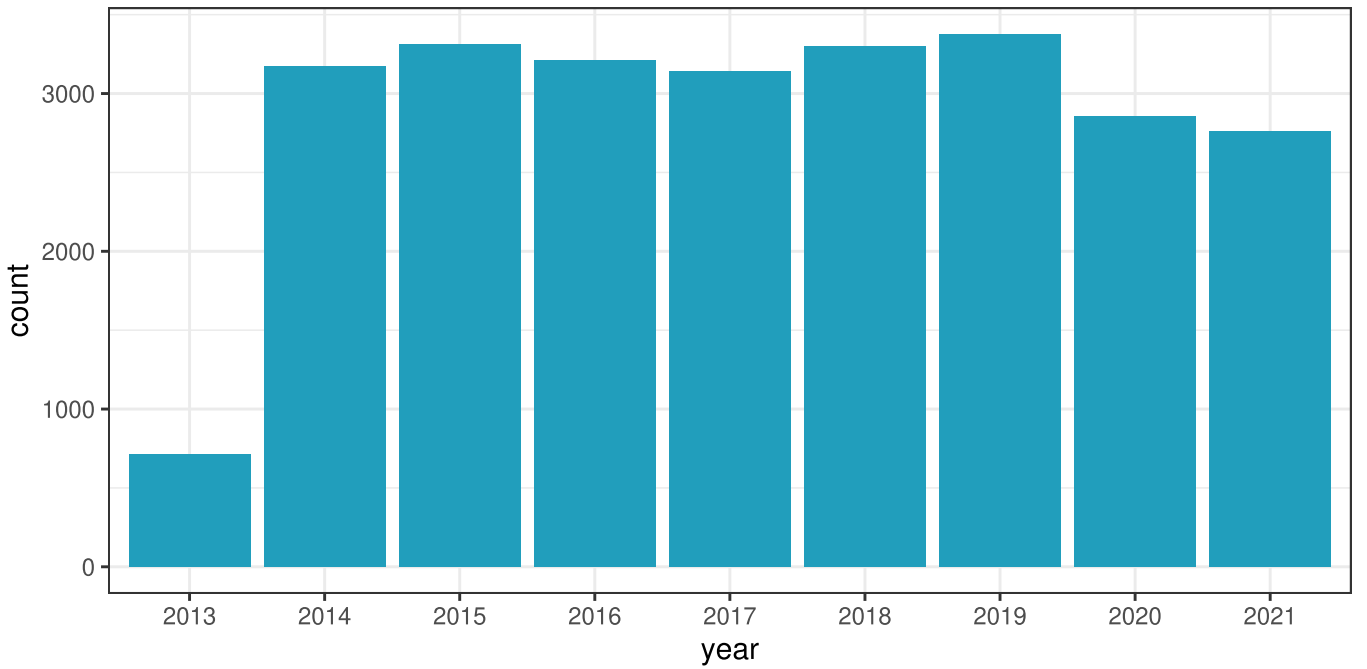
```
# A tibble: 2 × 14
  year  type  place weight_lbs grower_name     city  state_prov country gpc_site
  <chr> <chr> <chr>      <dbl> <chr>           <chr> <chr>      <chr>   <chr>
1 2013  S     1           1264 Pierpont, Edwin Jeff… Maine      United… Damaris…
2 2013  S     2           1233 Hunt, Jane & P… Came… Ontario    Canada  Port El…
# … with 5 more variables: seed_mother <chr>, pollinator_father <chr>,
#   ott <dbl>, est_weight <dbl>, pct_chart <dbl>
```

```
count(pumpS, year)
```

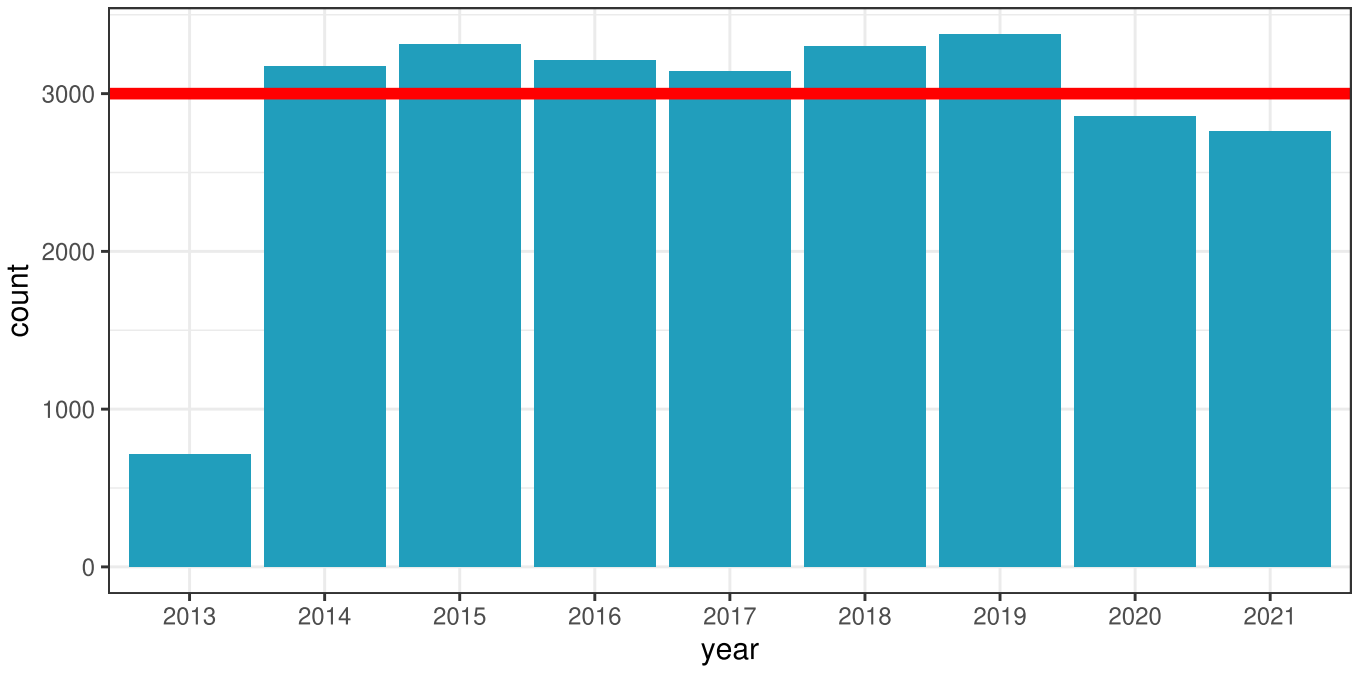
```
# A tibble: 9 × 2
  year      n
  <chr> <int>
1 2013    713
2 2014   3173
3 2015   3313
4 2016   3211
5 2017   3139
6 2018   3297
7 2019   3375
8 2020   2854
9 2021   2758
```

Исследование групп

```
ggplot(pumpS) +
  geom_bar(aes(x = year), fill = '#219ebc') + theme_bw()
```



```
ggplot(pumpS) +
  geom_bar(aes(x = year), fill = '#219ebc') +
  geom_hline(yintercept=3000, color = 'red', size = 2) +
  theme_bw()
```



Добавляем характеристики

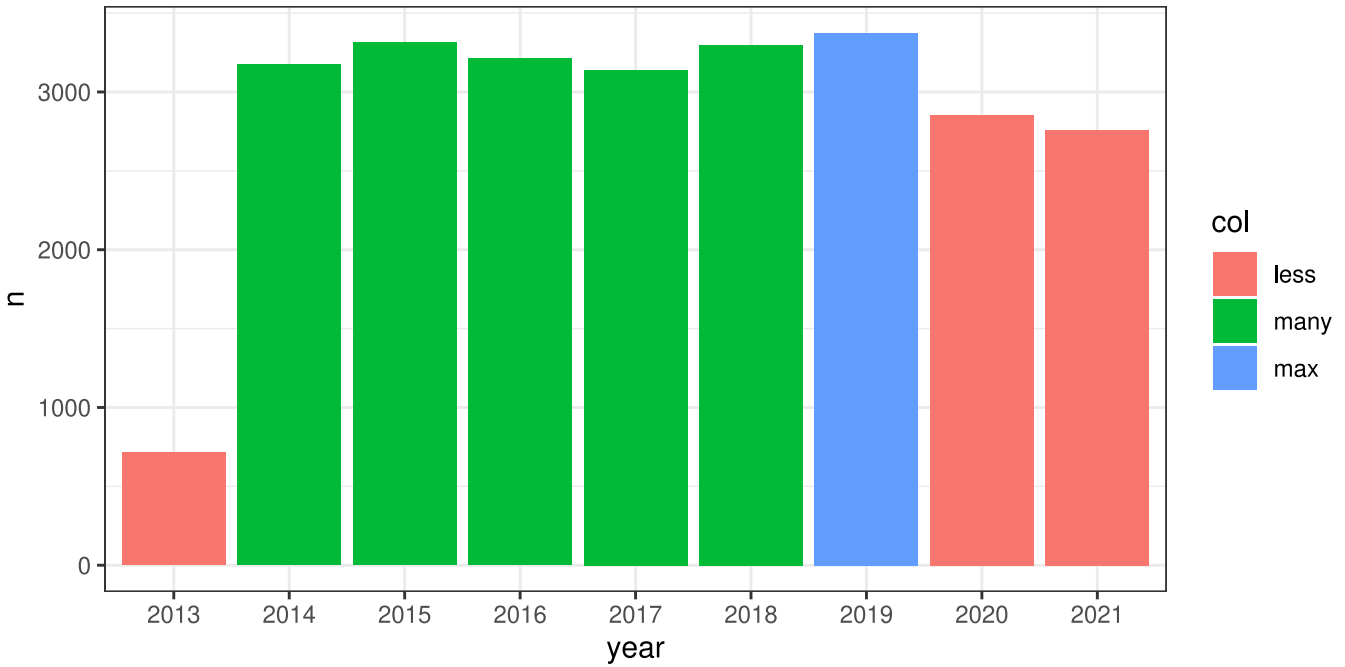
Можно организовать метаданные

```
pumpS %>%
  dplyr::count(year) %>%
  mutate(col = case_when(n == max(n) ~ 'max',
                         n > 3000 ~ 'many',
                         n <= 3000 ~ 'less')) -> pump_year
pump_year
```

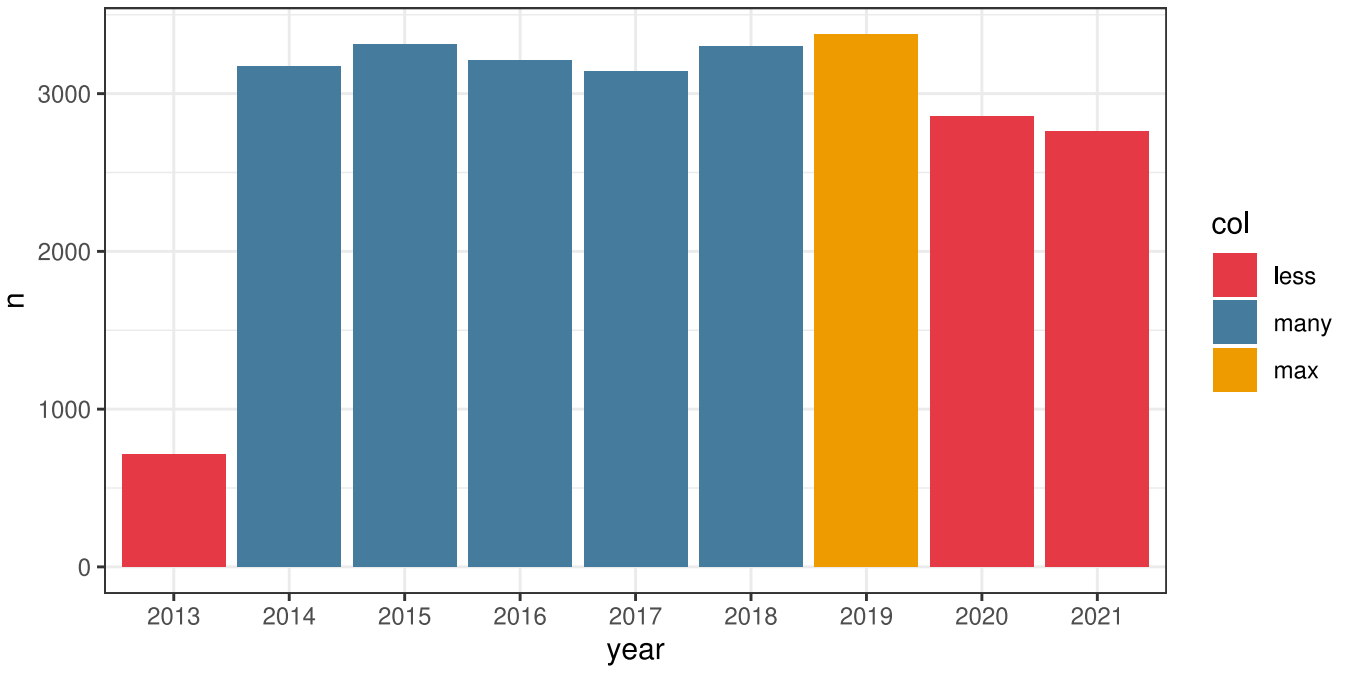
```
# A tibble: 9 × 3
  year      n col
  <chr> <int> <chr>
1 2013    713 less
2 2014   3173 many
3 2015   3313 many
4 2016   3211 many
5 2017   3139 many
6 2018   3297 many
7 2019   3375 max
8 2020   2854 less
9 2021   2758 less
```

Цветом можно добавлять больше информации

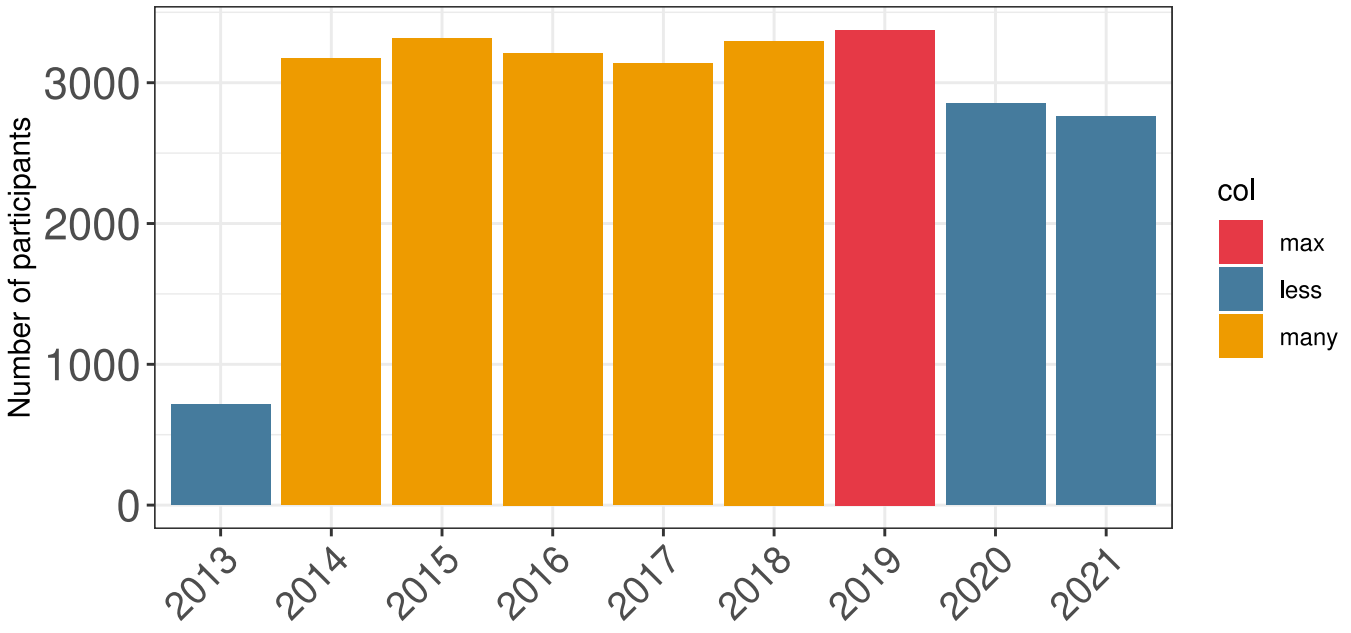
```
pump_year %>%
  ggplot() + geom_bar(aes(x = year, y = n, fill = col), stat = 'identity') +
  theme_bw()
```



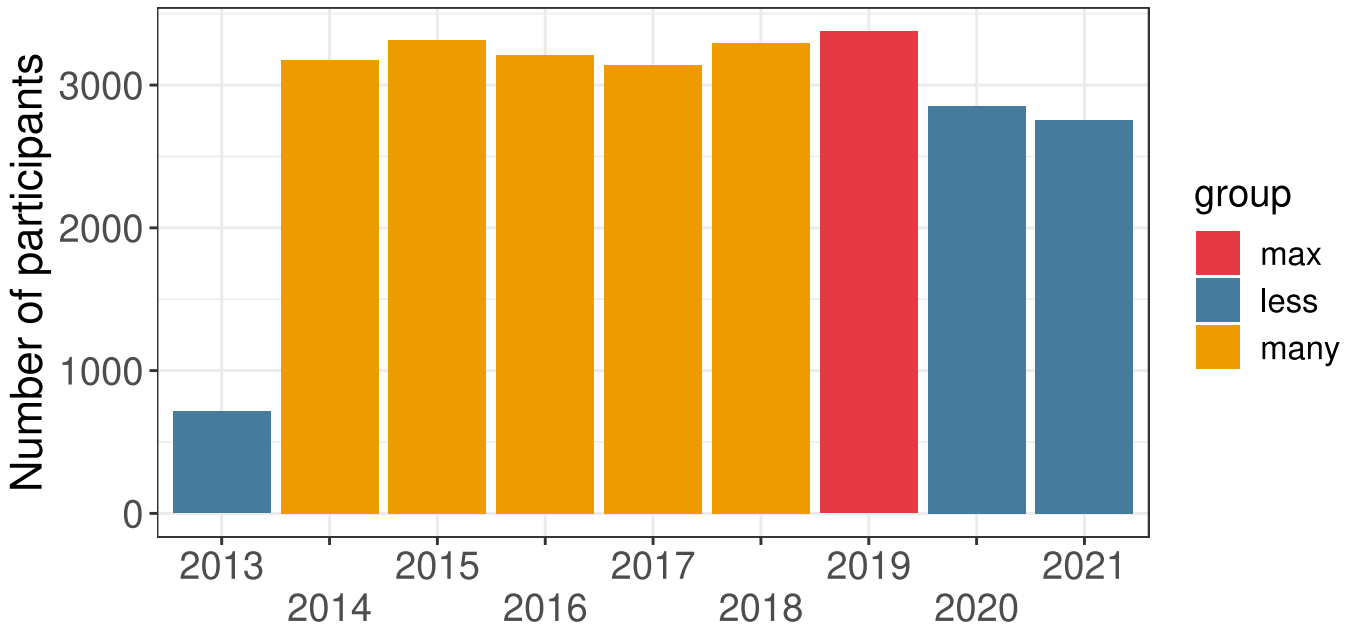
```
pump_year %>%
  ggplot() + geom_bar(aes(x = year, y = n, fill = col), stat = 'identity') +
  theme_bw() +
  scale_fill_manual(values = c('#e63946','#457b9d','#ee9b00'))
```



```
pump_year %>%
  ggplot() + geom_bar(aes(x = year, y = n, fill = col), stat = 'identity') +
  theme_bw() +
  scale_fill_manual(values = c(max = '#e63946',less = '#457b9d', many = '#ee9b00')) +
  theme(axis.text.x = element_text(size = 14, angle = 45, hjust = 1),
        axis.text.y = element_text(size = 16)) +
  labs(x = "", y = "Number of participants")
```

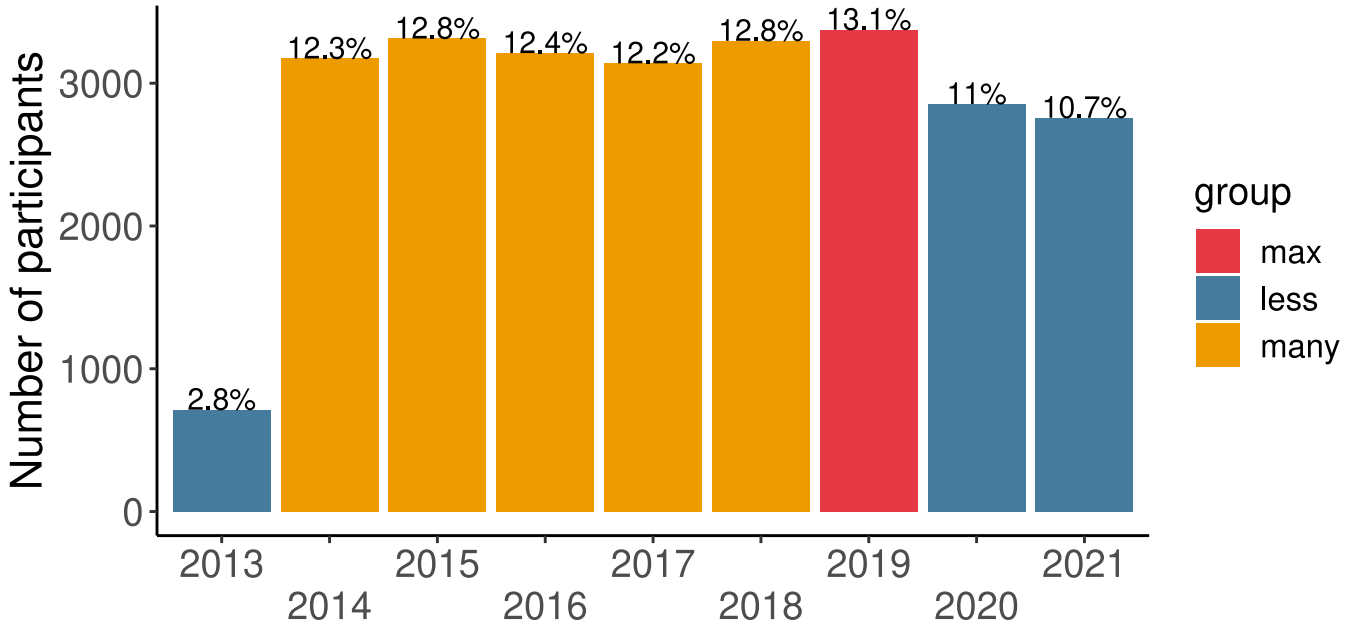


```
pump_year %>%
  ggplot() + geom_bar(aes(x = year, y = n, fill = col), stat = 'identity') +
  theme_bw() +
  scale_fill_manual(values = c(max = '#e63946',less = '#457b9d', many = '#ee9b00')) +
  scale_x_discrete(guide = guide_axis(n.dodge = 2)) +
  theme(axis.text = element_text(size = 14),
        axis.title.y = element_text(size = 16),
        legend.text = element_text(size=12),
        legend.title = element_text(size=14)) +
  labs(x = "", y = "Number of participants", fill = "group")
```

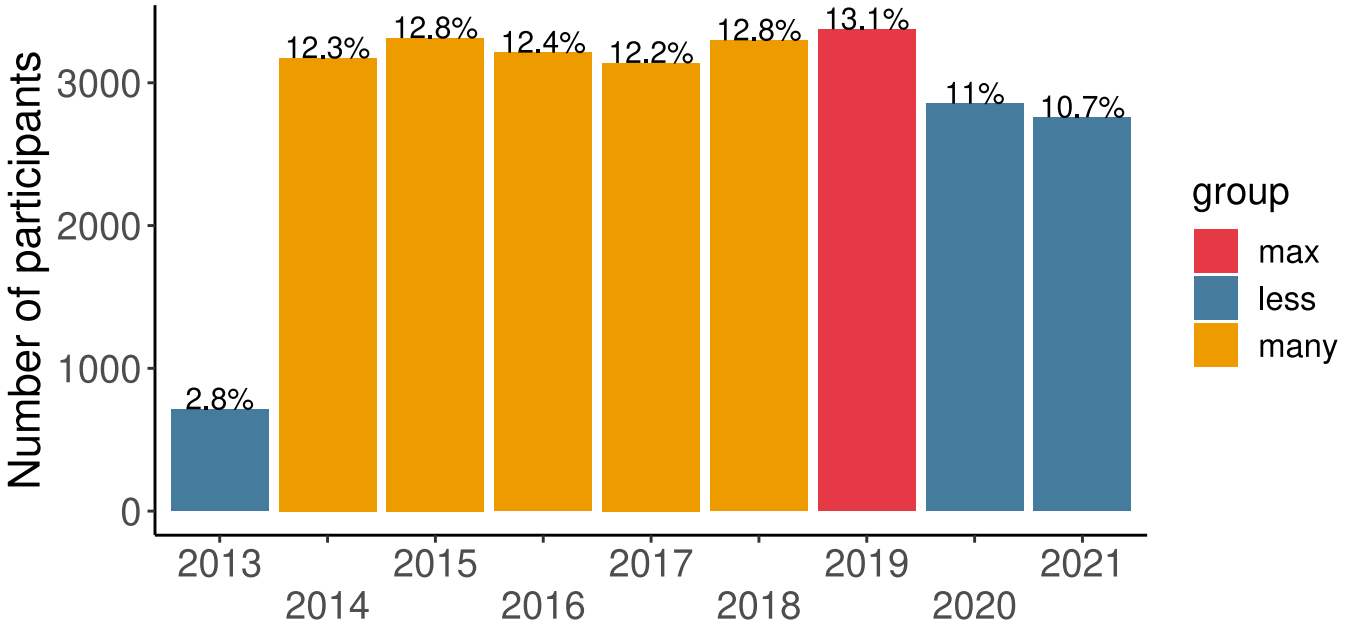


```
goodsize <-
    ggplot2::theme_classic() +
    ggplot2::theme(axis.text = element_text(size = 14),
        axis.title.y = element_text(size = 16),
        legend.text = element_text(size=12),
        legend.title = element_text(size=14))
```

```
theme_set(goodsize)
pump_year %>%
  mutate(per = round((n / sum(n))*100,1)) %>%
  ggplot(aes(x = year, y = n, fill = col)) + geom_bar(stat = 'identity') +
  scale_fill_manual(values = c(max='#e63946',less='#457b9d',many='#ee9b00')) +
  scale_x_discrete(guide = guide_axis(n.dodge = 2)) +
  labs(x = "", y = "Number of participants", fill = "group") +
  geom_text(aes(label=paste(per,'%',sep='')), vjust=0)
```

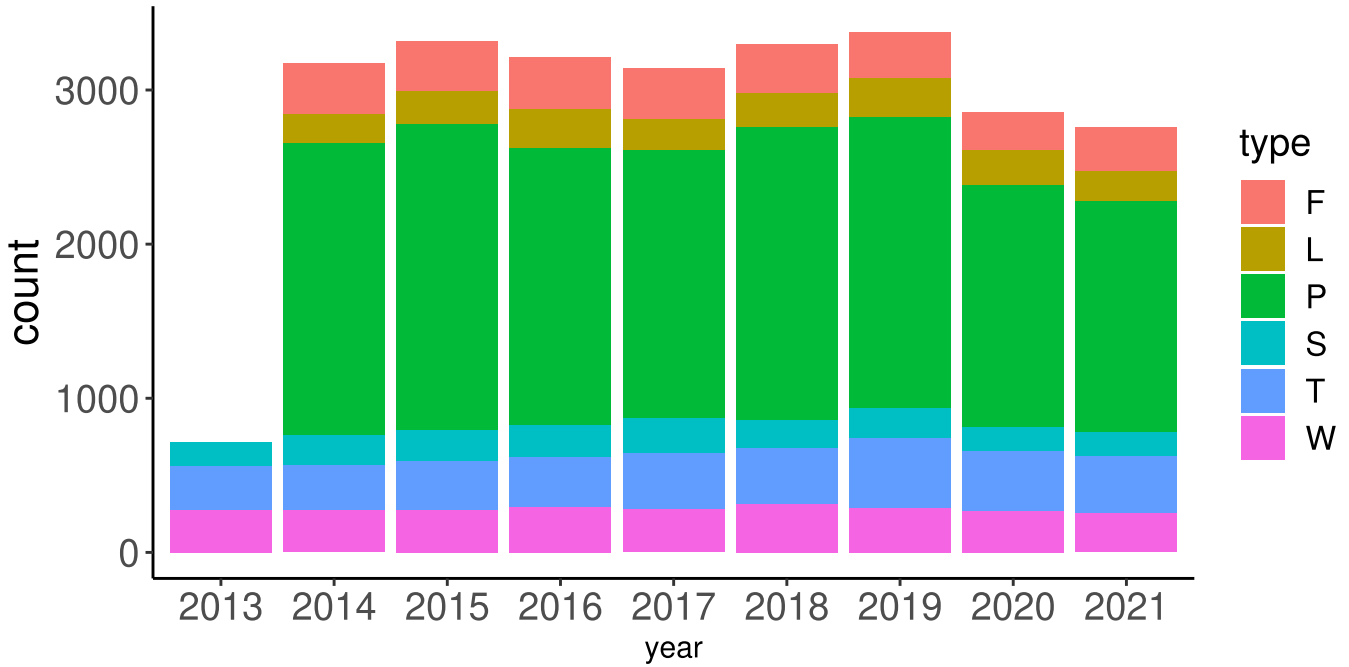


```
pump_year %>%
  mutate(per = round((n / sum(n))*100,1)) %>%
  ggplot(aes(x = year, y = n, fill = col)) + geom_bar(stat = 'identity') +
  scale_fill_manual(values = c(max='#e63946',less='#457b9d',many='#ee9b00')) +
  scale_x_discrete(guide = guide_axis(n.dodge = 2)) +
  labs(x = "", y = "Number of participants", fill = "group") +
  geom_text(aes(label=paste(per,'%',sep='')), vjust=0)
```



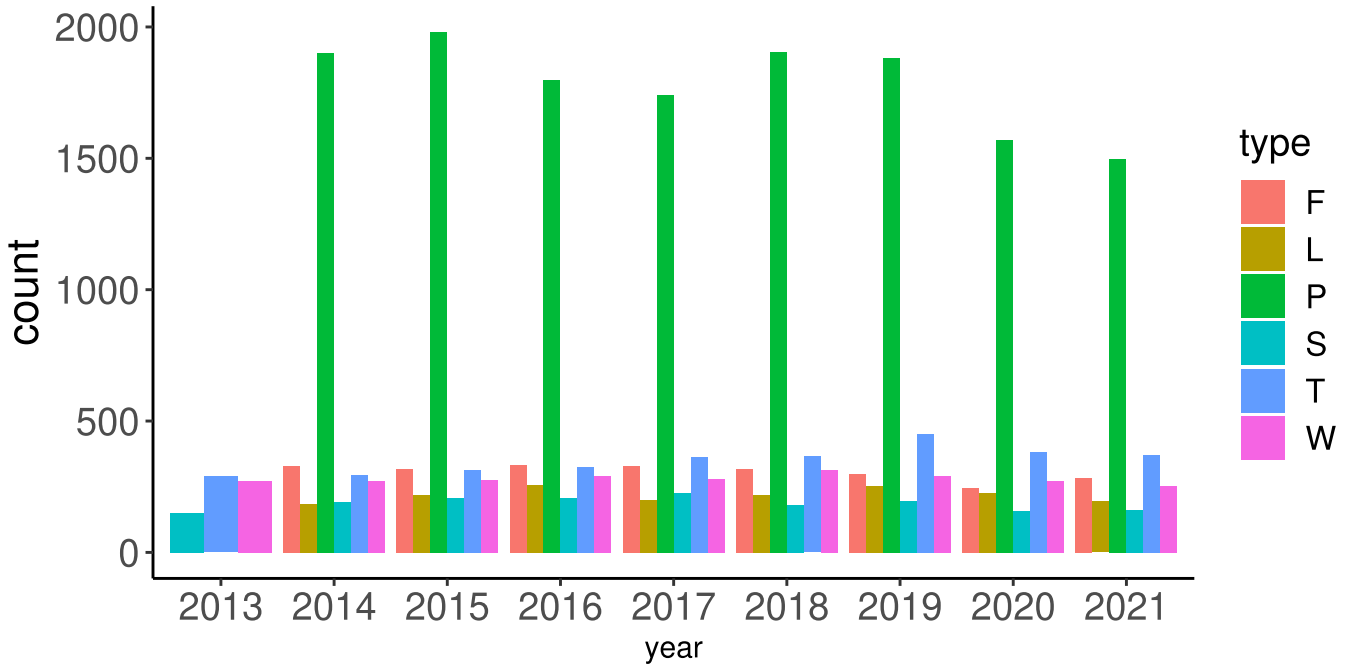
Изучение структуры групп

```
pumpS %>%
  ggplot() +
  geom_bar(aes(x = year, fill = type))
```

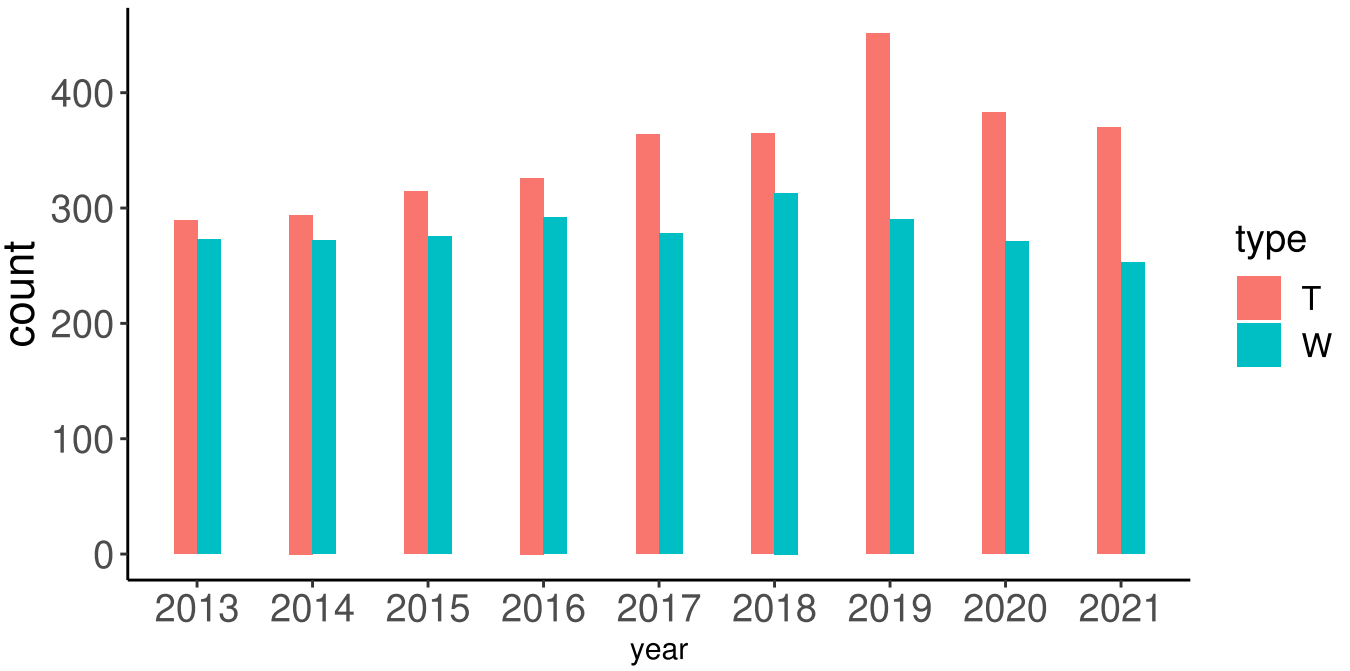


Много групп + много подргупп = плохо

```
pumpS %>%
  ggplot() +
  geom_bar(aes(x = year, fill = type), position=position_dodge())
```

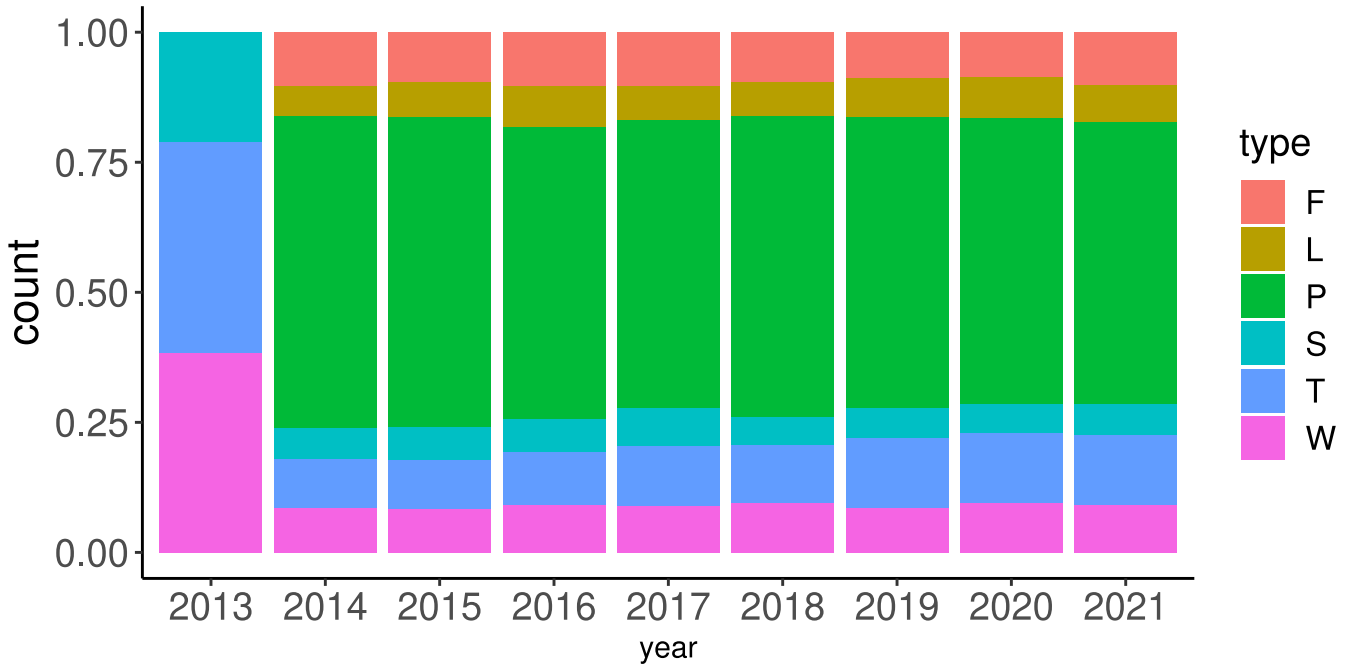


```
pumpS %>%
  filter(type %in% c('W','T')) %>%
  ggplot() +
  geom_bar(aes(x = year, fill = type), position=position_dodge(), width = 0.4)
```

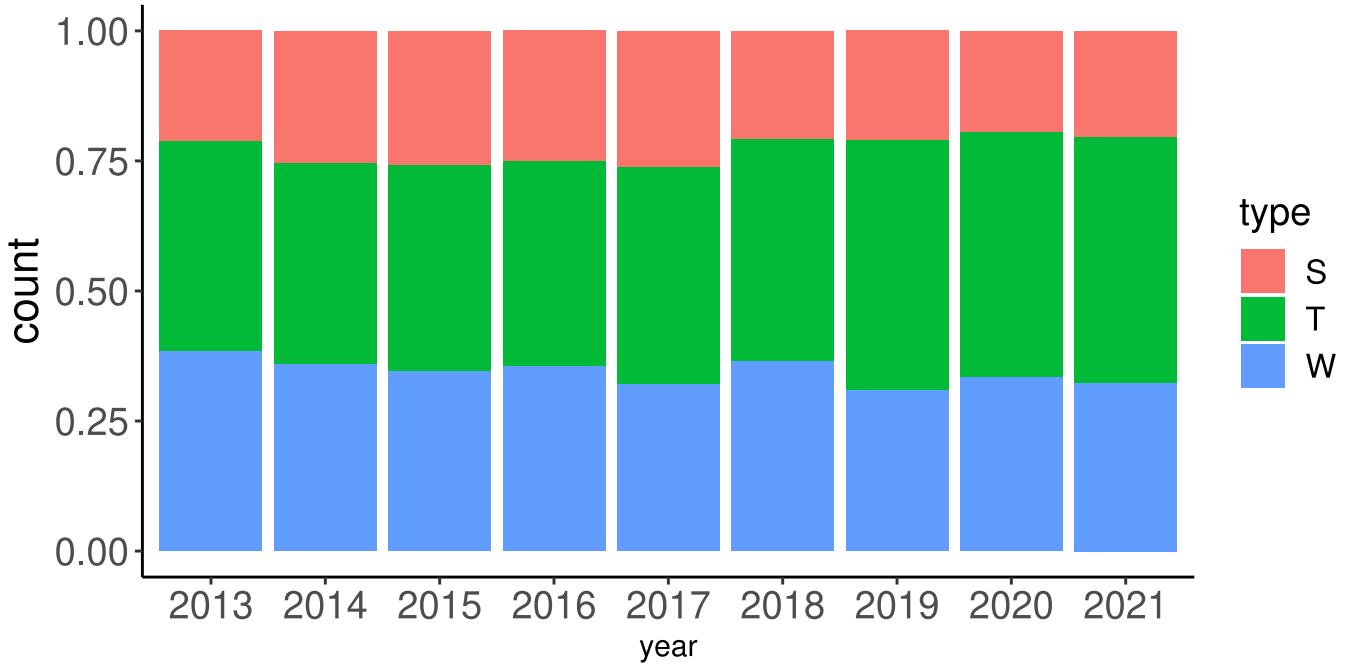


Абсолютные значения vs проценты

```
pumpS %>%
  ggplot() +
  geom_bar(aes(x = year, fill = type), position = 'fill')
```

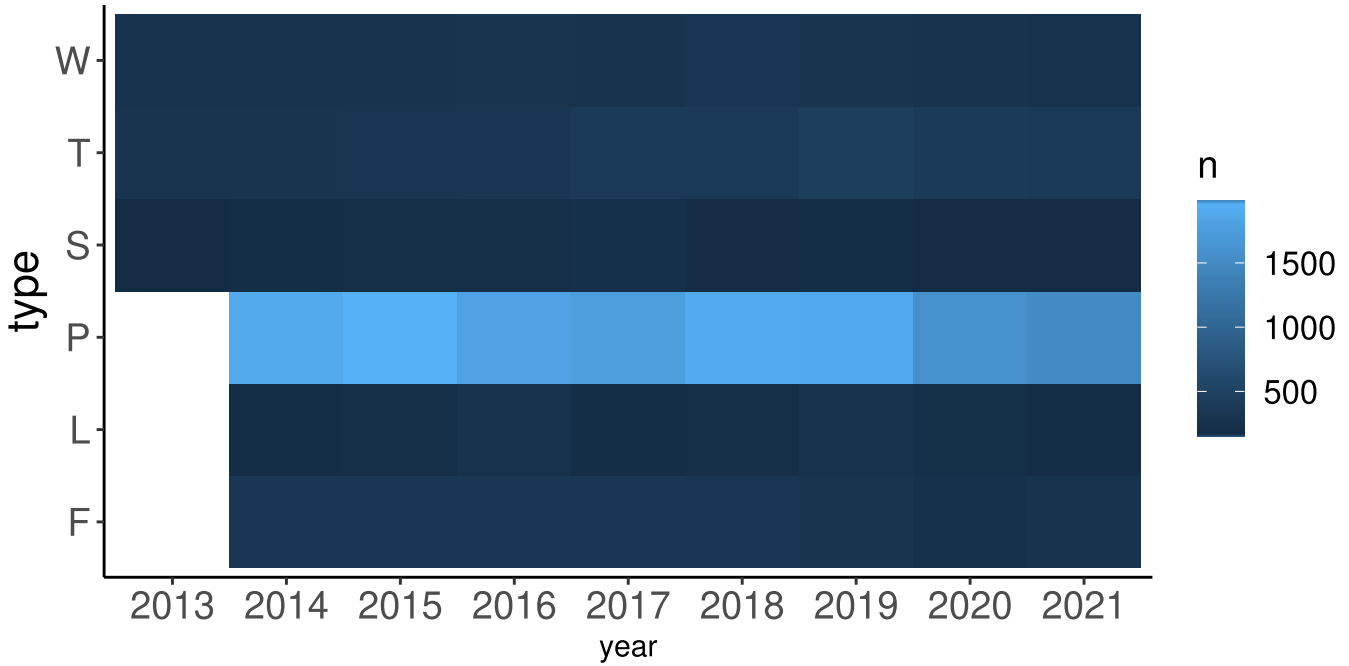


```
pumpS %>%
  filter(type %in% c('W', 'T', 'S')) %>%
  ggplot() +
  geom_bar(aes(x = year, fill = type), position = 'fill')
```

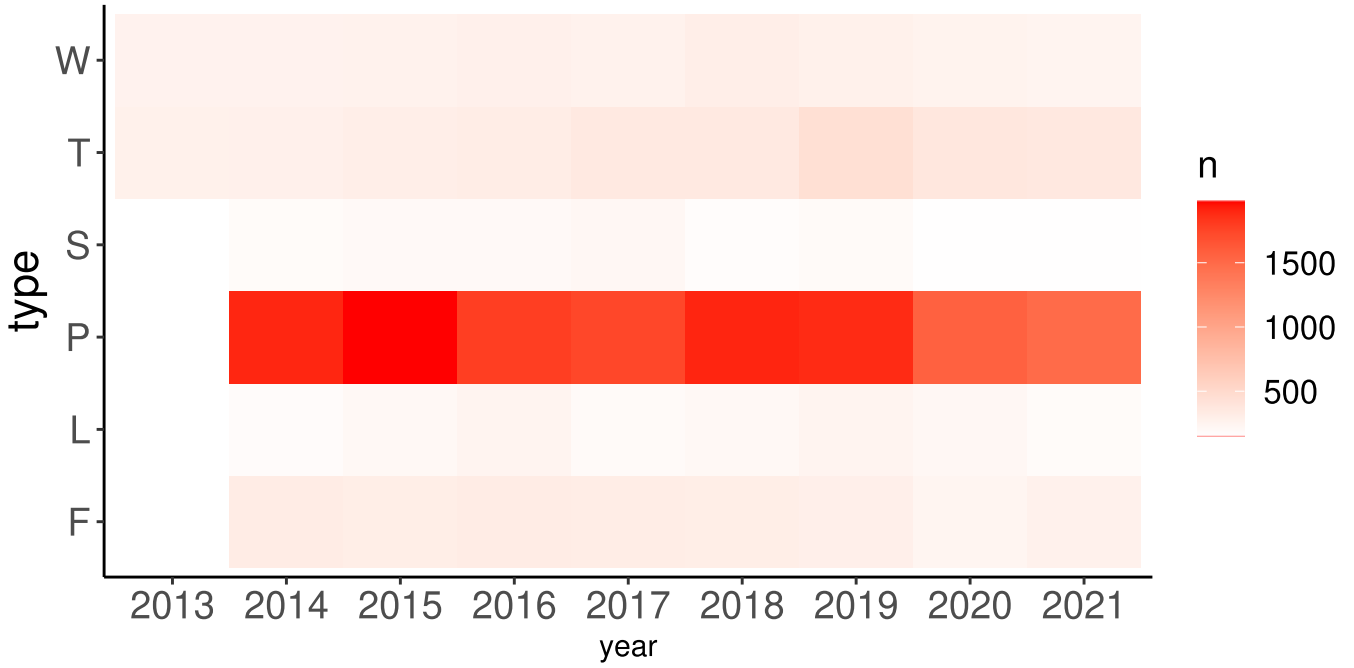


Тепловые карты

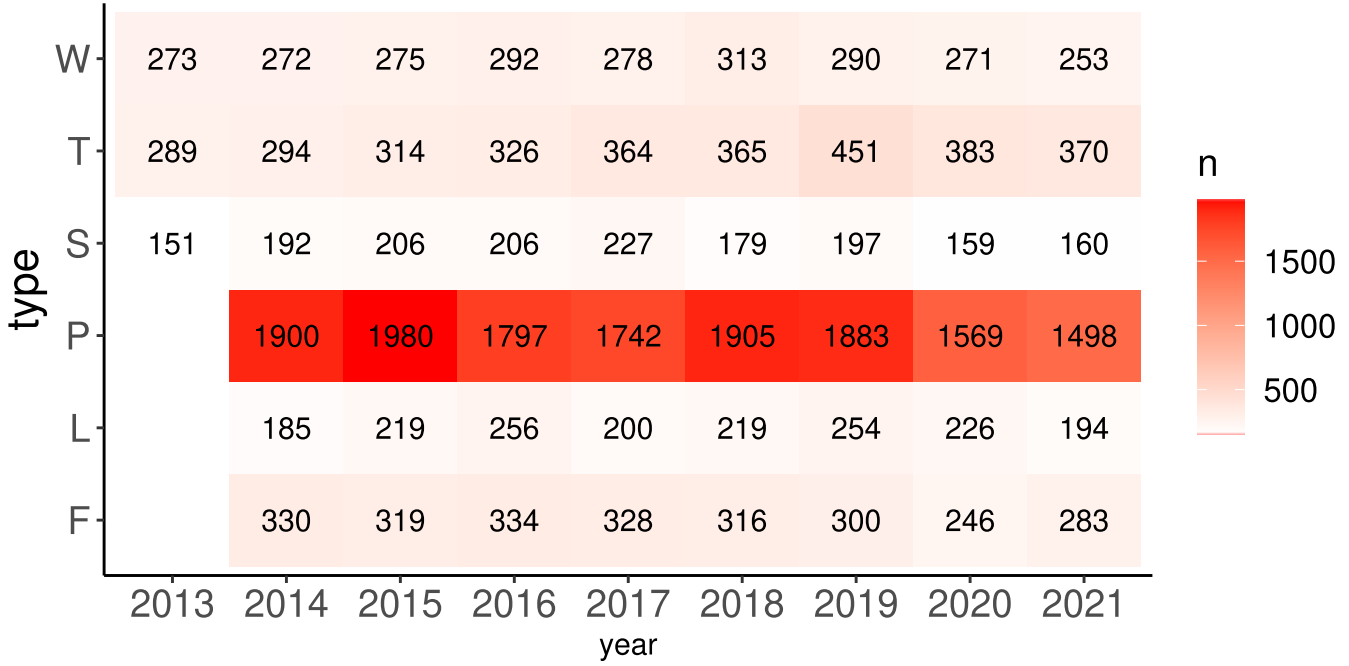
```
pumpS %>%
  dplyr::count(year, type) -> pumpYT

pumpYT %>%
  ggplot(aes(x = year, y = type, fill = n)) +
  geom_tile()
```



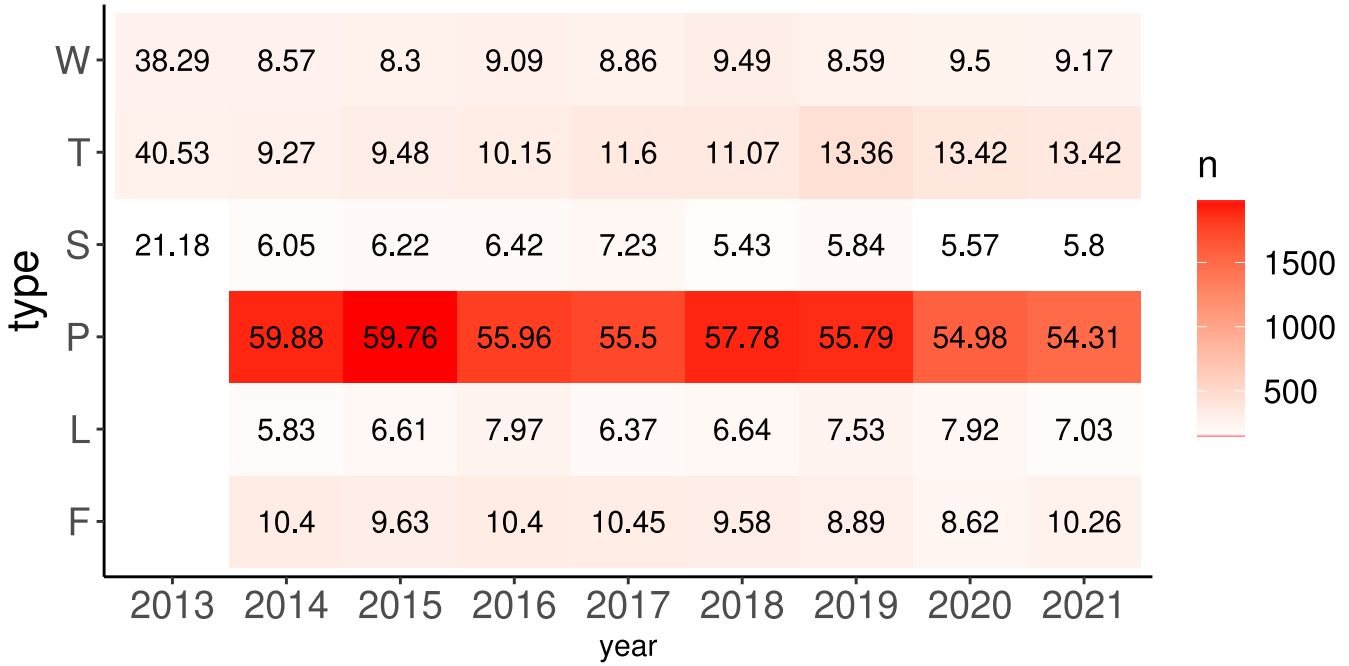
```
pumpYT %>%
  ggplot(aes(x = year, y = type, fill = n)) +
  geom_tile() +
  scale_fill_gradient(low = "white", high = "red")
```



```
pumpYT %>%
  ggplot(aes(x = year, y = type, fill = n)) +
  geom_tile() +
  geom_text(aes(label = n)) +
  scale_fill_gradient(low = "white", high = "red")
```



```
pumpYT %>%
  group_by(year) %>%
  mutate(per = round((n / sum(n))*100,2)) %>%
  ungroup() %>%
  ggplot(aes(x = year, y = type, fill = n)) +
  geom_tile() +
  geom_text(aes(label = per)) +
  scale_fill_gradient(low = "white", high = "red")
```



Полное попарное сочетание значений из двух столбцов

```
tr <- tibble(
  a = c('a', 'b' ,'b'),
  b = c('q', 'w', 'e'))
tr
```

```
# A tibble: 3 × 2
  a     b
  <chr> <chr>
1 a     q
2 b     w
3 b     e
```

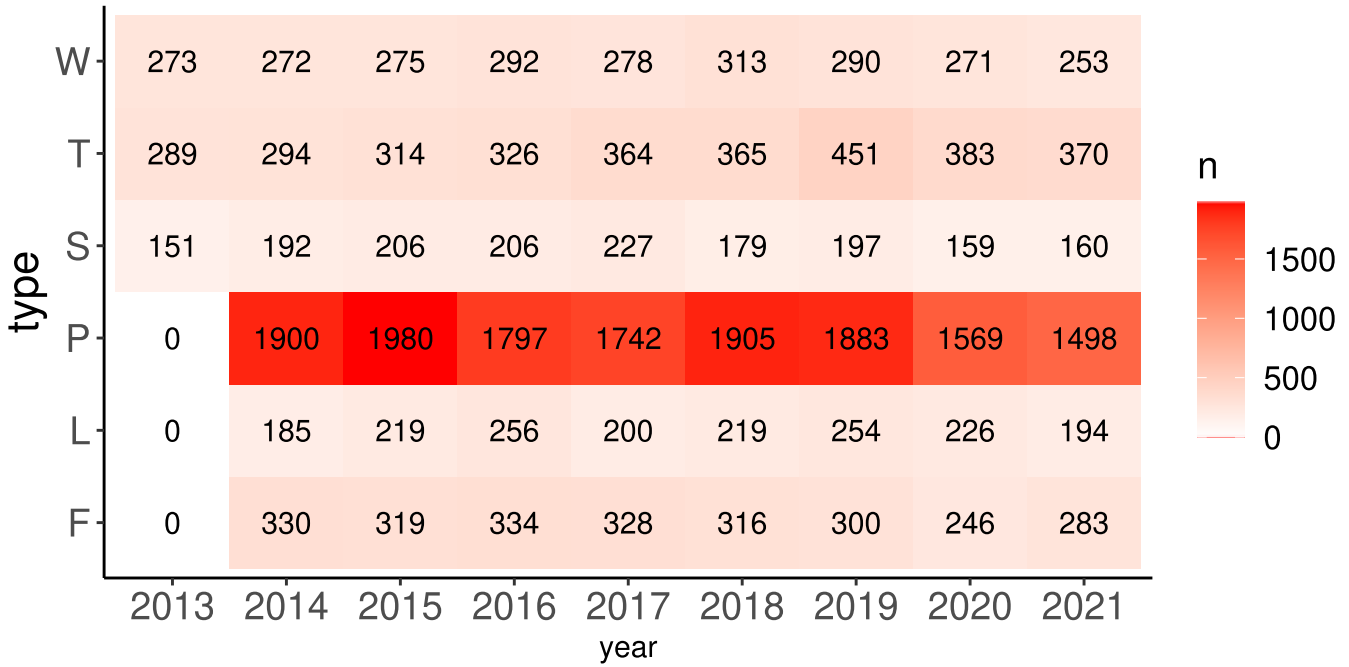
```
tidyr::expand(tr, a, b)
```

```
# A tibble: 6 × 2
  a     b
  <chr> <chr>
1 a     e
2 a     q
3 a     w
4 b     e
5 b     q
6 b     w
```

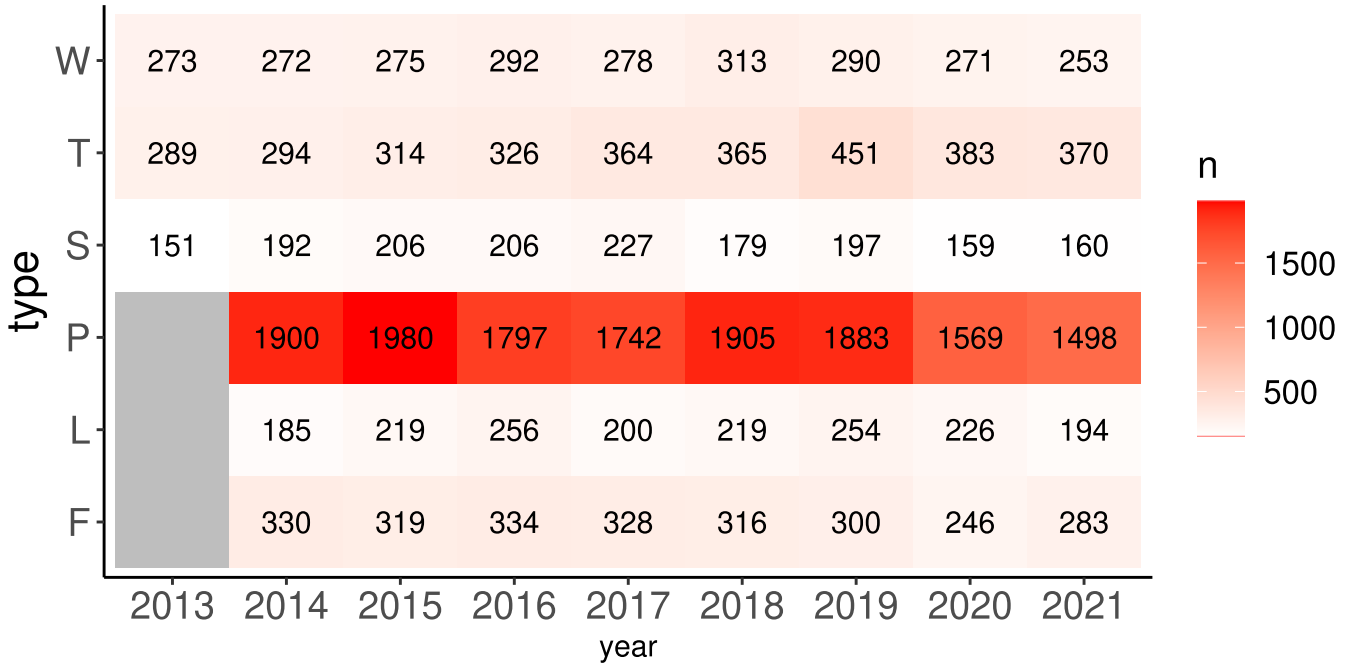
Замена NA

можно на 0, но нужно ли?

```
pumpYT %>%
  tidyr::expand(year, type) %>%
  left_join(pumpYT, by = c("year", "type")) %>%
  mutate( n = ifelse(is.na(n), 0, n)) %>%
  ggplot(aes(x = year, y = type, fill = n)) + geom_tile() +
  geom_text(aes(label = n)) +
  scale_fill_gradient(low = "white", high = "red")
```

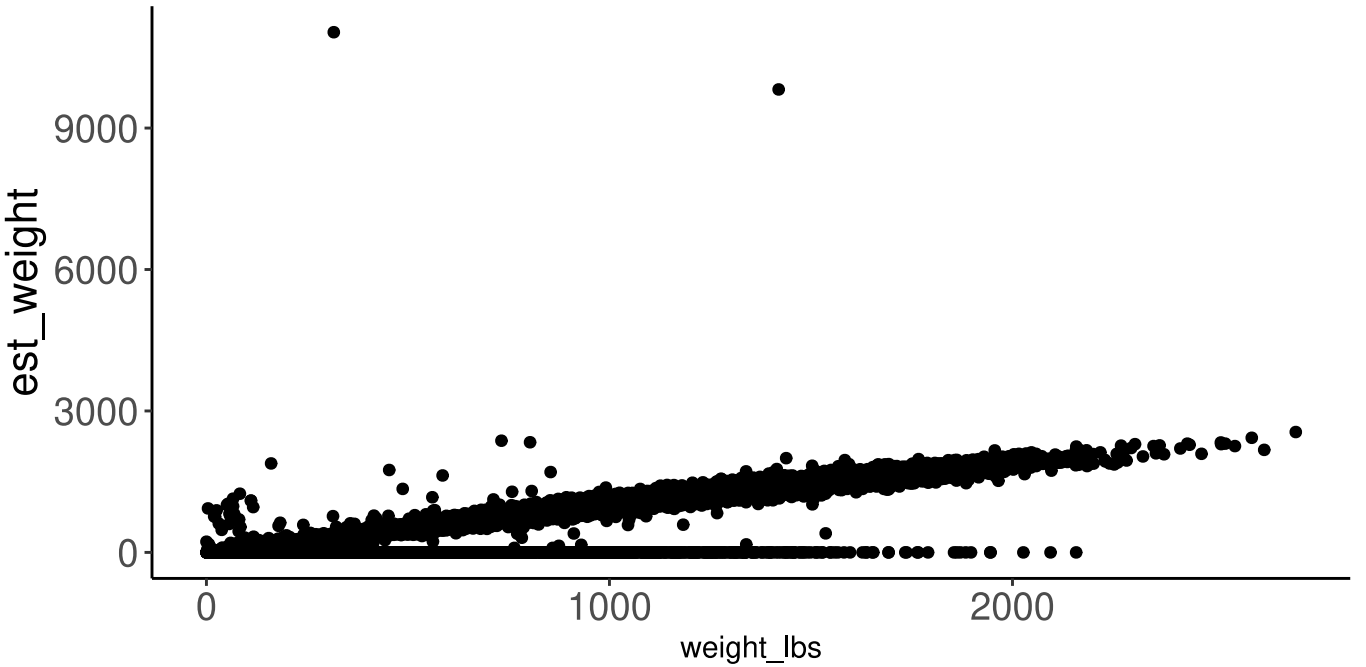


```
pumpYT %>%
  tidyr::expand(year, type) %>%
  left_join(pumpYT, by = c("year", "type")) %>%
  ggplot(aes(x = year, y = type, fill = n)) +
  geom_tile() +
  scale_fill_gradient(low = "white", high = "red", na.value = 'grey') +
  geom_text(aes(label = n))
```

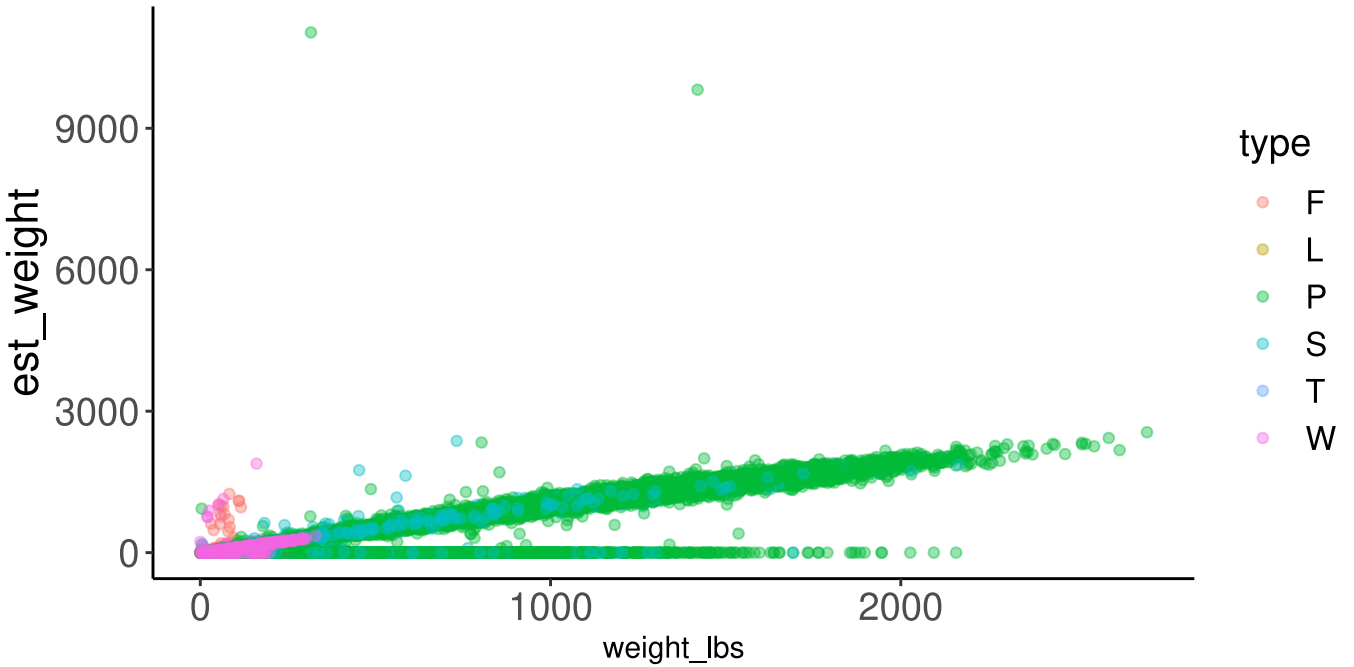


Выбросы?

```
pumpS %>%
  ggplot() +
  geom_point(aes(x = weight_lbs, y = est_weight))
```



```
pumpS %>%
  ggplot() +
  geom_point(aes(x = weight_lbs, y = est_weight, color = type), alpha = 0.4)
```



Столбец type

F - field pumpkin

P - giant pumpkin

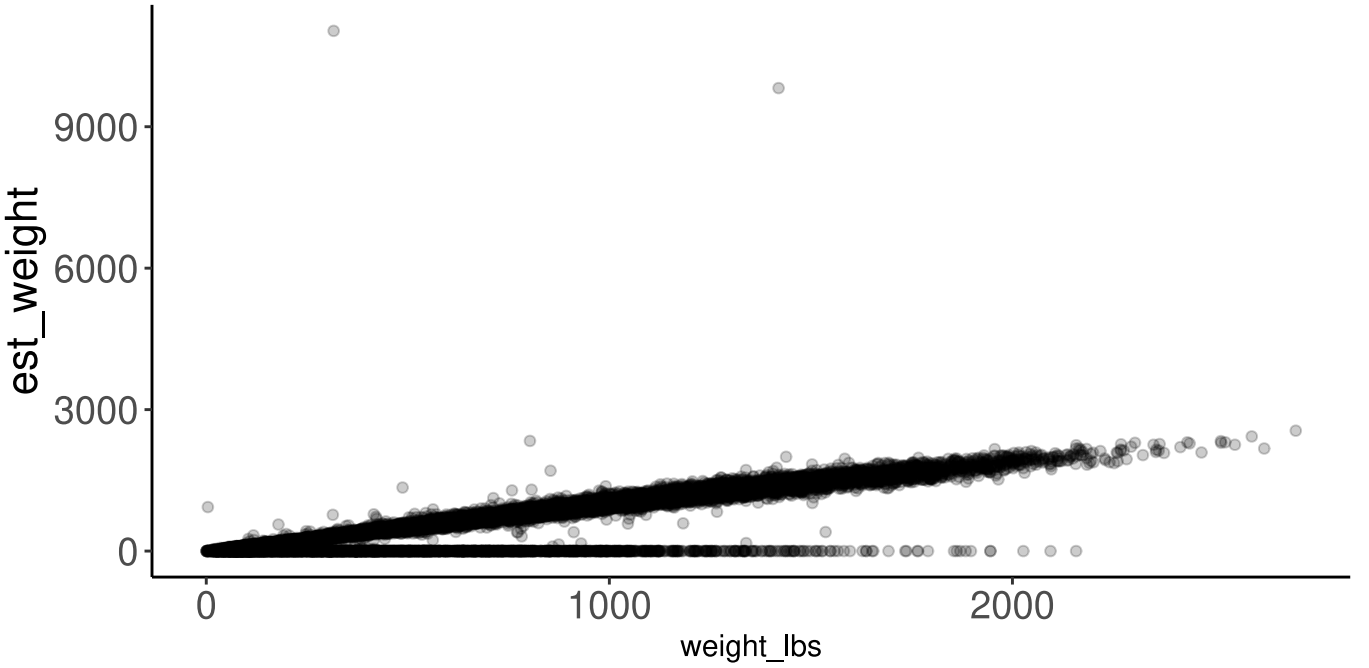
S - giant squash

W - giant watermelon

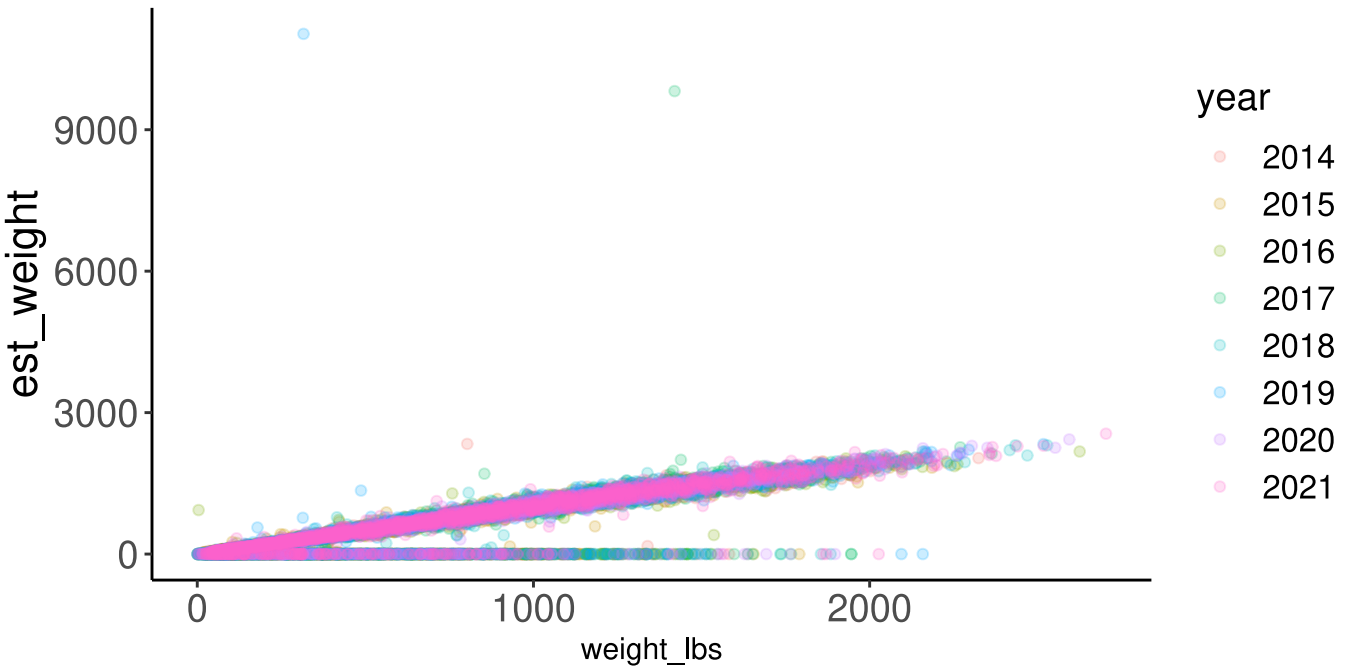
L - long gourd

T - tomato

```
pumpS %>%
  filter(type == 'P') %>%
  ggplot() +
  geom_point(aes(x = weight_lbs, y = est_weight), alpha = 0.2)
```



```
pumpS %>%
  filter(type == 'P') %>%
  ggplot() +
  geom_point(aes(x = weight_lbs, y = est_weight, color = year), alpha = 0.2)
```



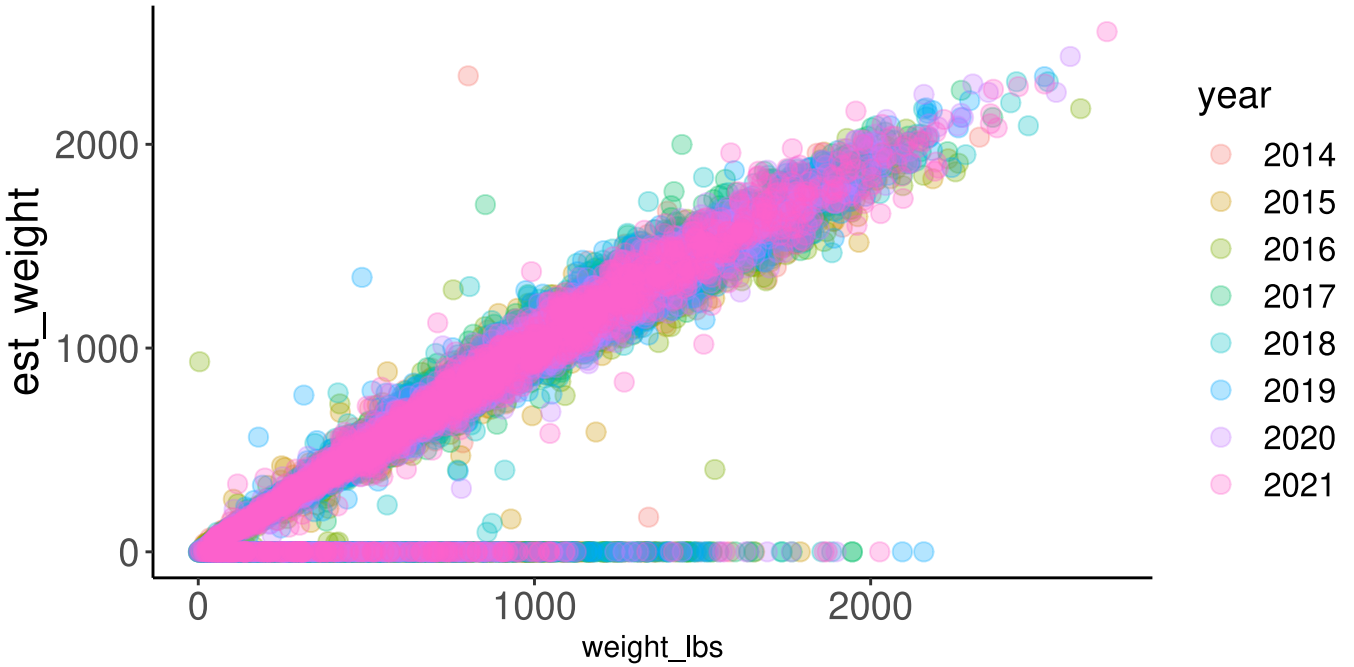
На выбросы стоит посмотреть отдельно

```
pumpS %>%
  filter(type == 'P', est_weight > 9000)
```

```
# A tibble: 2 × 14
  year  type  place weight_lbs grower_name     city  state_prov country gpc_site
  <chr> <chr> <chr>      <dbl> <chr>           <chr> <chr>      <chr>   <chr>
1 2017  P     215         1420. Wuersching, Ma… Einh… Hesse      Germany Early W…
2 2019  P     1282         316  Shenfish, Gary  Litt… Colorado   United… Fort Co…
# … with 5 more variables: seed_mother <chr>, pollinator_father <chr>,
#   ott <dbl>, est_weight <dbl>, pct_chart <dbl>
```

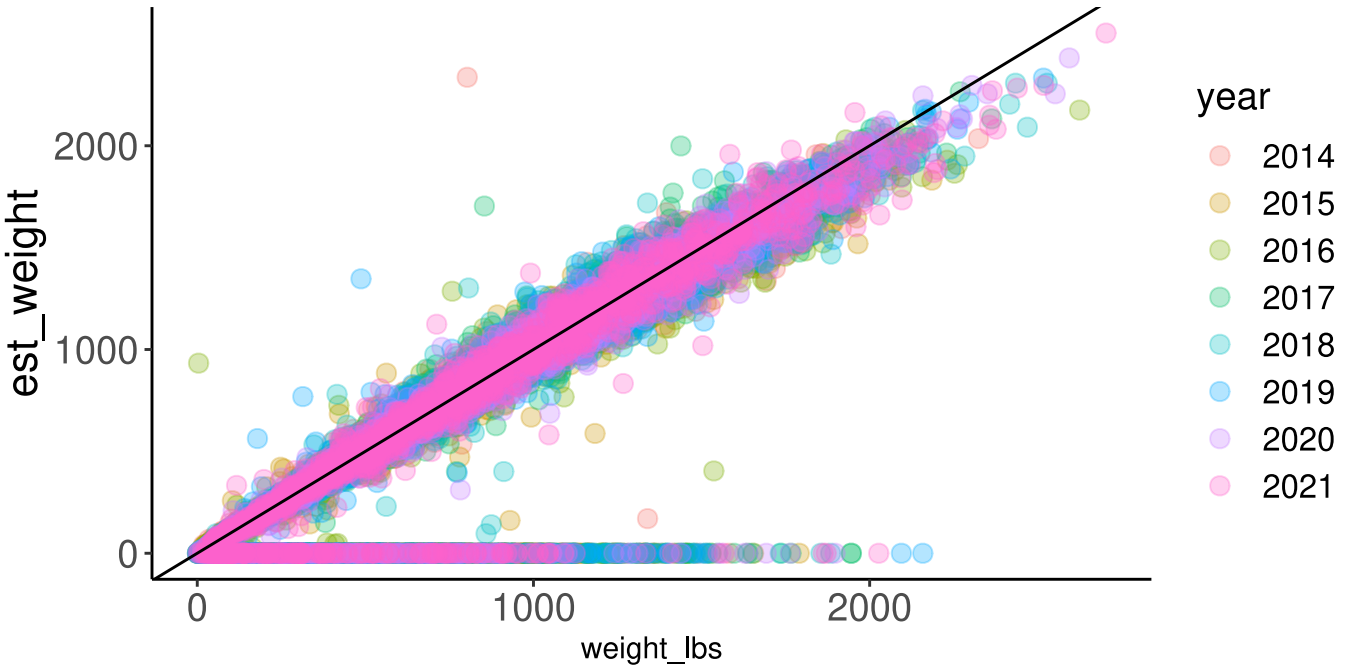
Опять выбросы?

```
pumpS %>%
  filter(type == 'P', est_weight < 9000) %>%
  ggplot() +
  geom_point(aes(x = weight_lbs, y = est_weight, color = year), alpha = 0.3, size = 3)
```



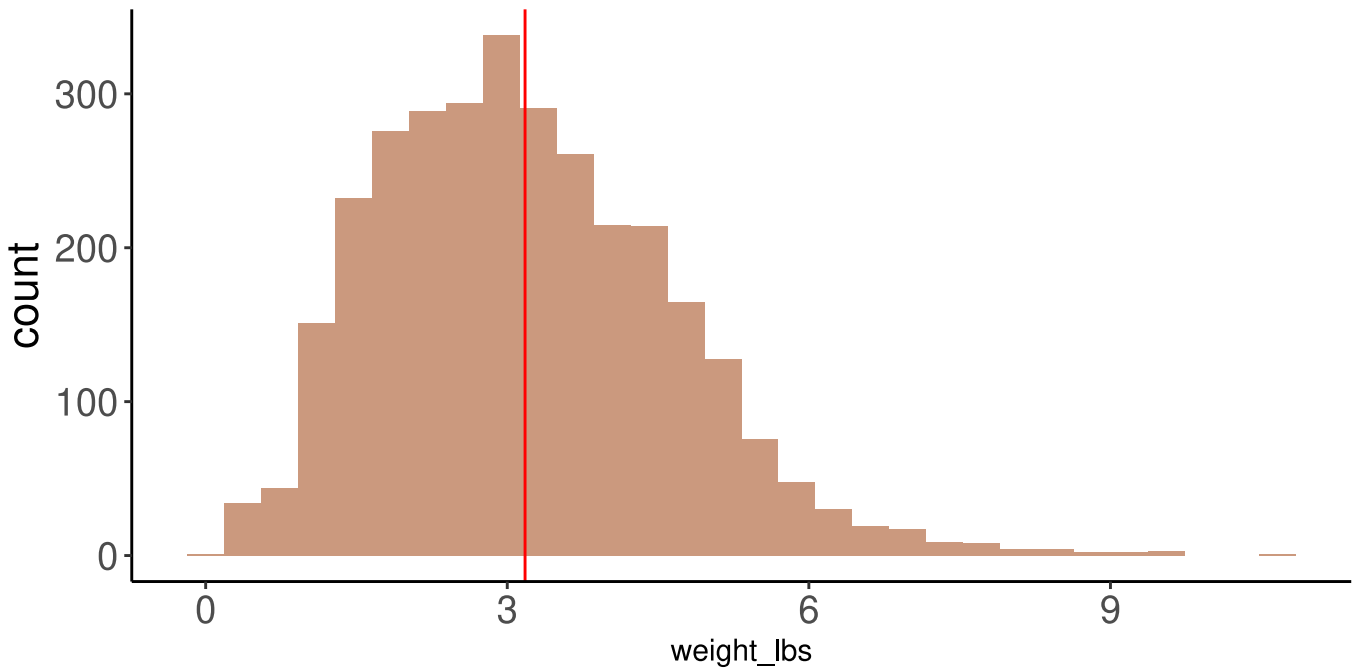
Опорные линии

```
pumpS %>%
  filter(type == 'P', est_weight < 9000) %>%
  ggplot() +
  geom_point(aes(x = weight_lbs, y = est_weight, color = year), alpha = 0.3, size = 3) +
  geom_abline(slope=1, intercept = 0)
```



Опорные линии

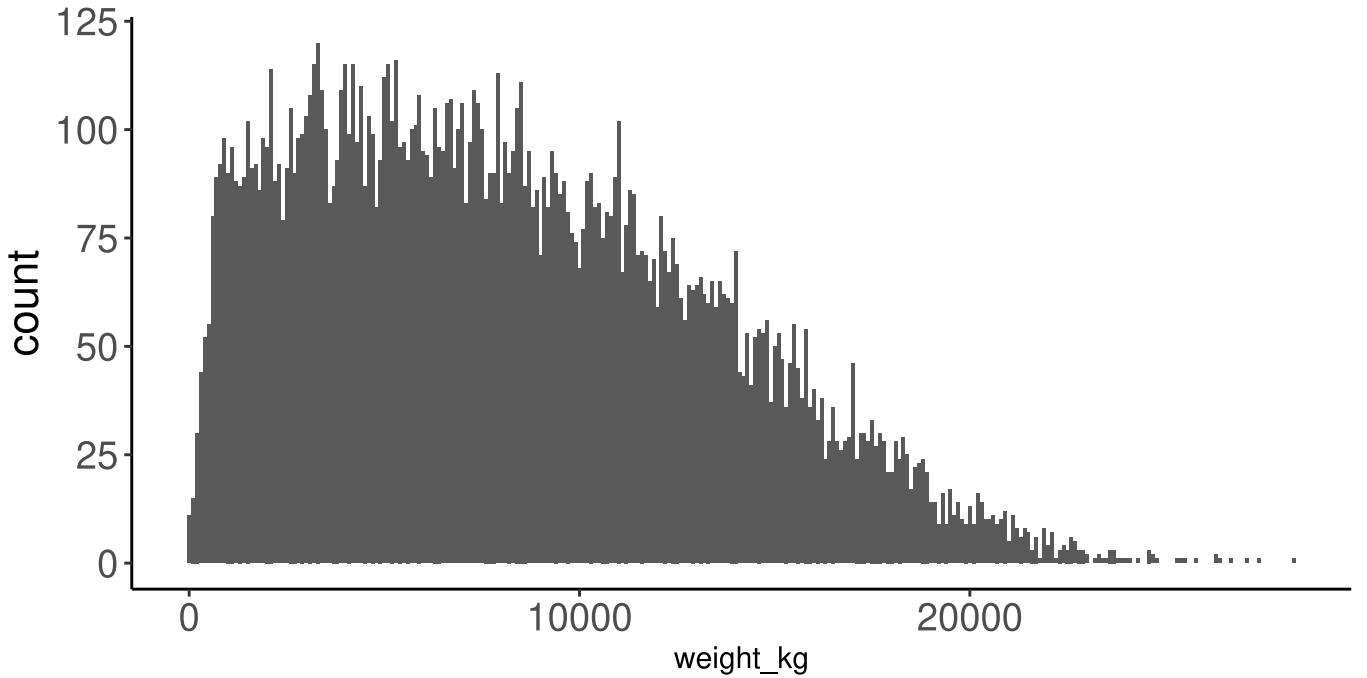
```
MT <- pumpS %>%
  filter(type == 'T') %>%
  summarise(MT = mean(weight_lbs))

pumpS %>%
  filter(type == 'T') %>%
  ggplot() +
  geom_histogram(aes(weight_lbs), fill = '#cb997e') +
  geom_vline(xintercept = MT$MT, col = 'red')
```



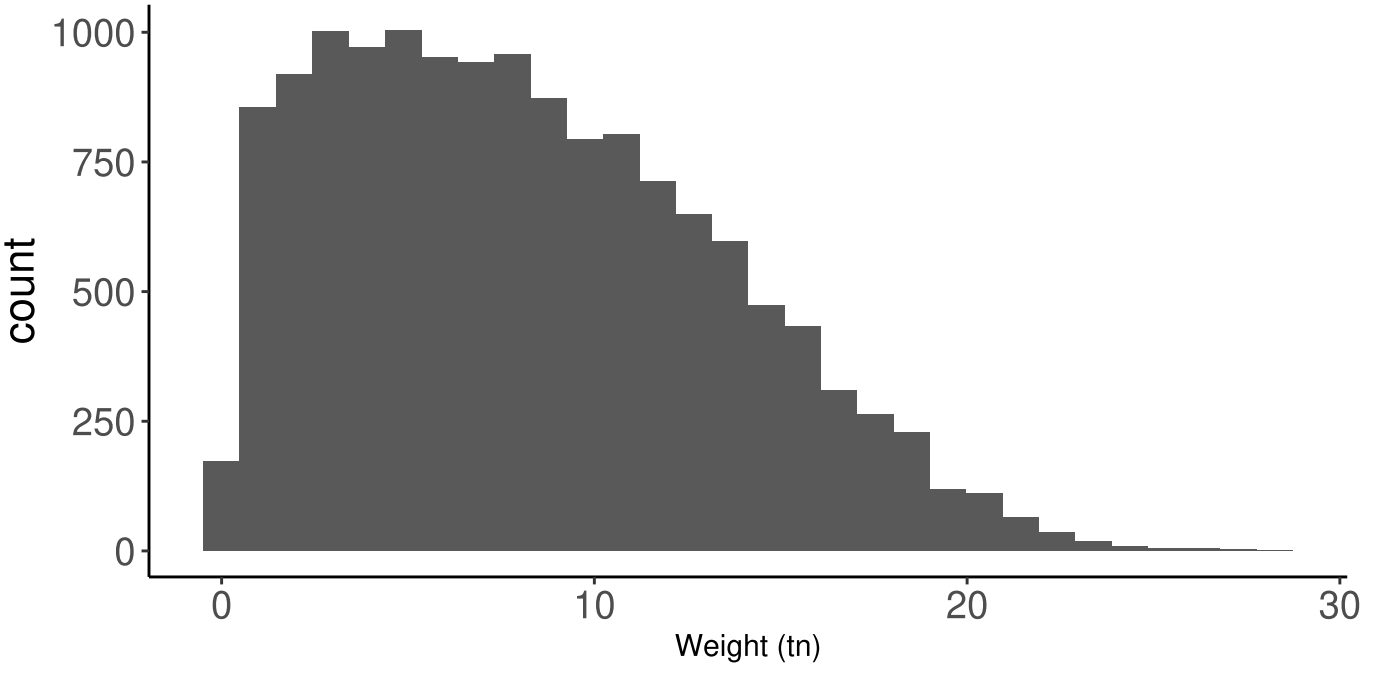
20 тонн???

Априорные знания о предмете

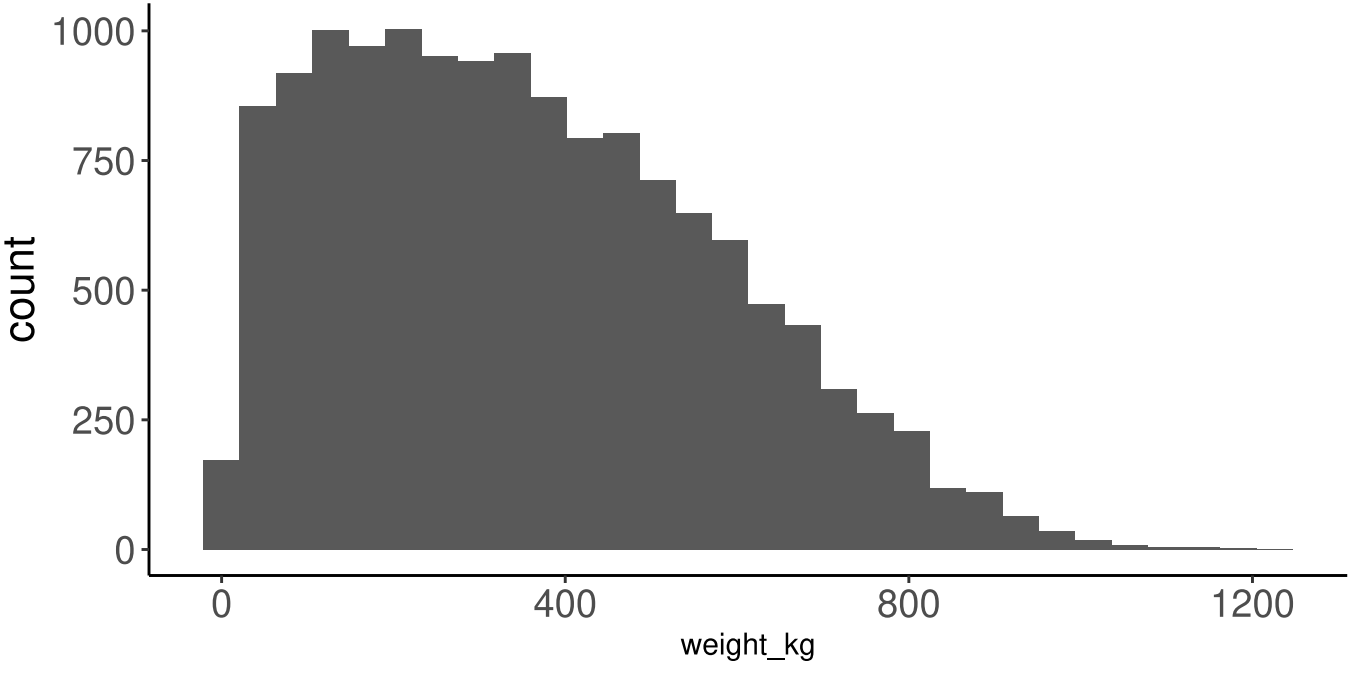
```
pumpS %>%
  filter(type == 'P', est_weight < 9000) %>%
  mutate(weight_kg = weight_lbs * 10.453592) %>%
  ggplot() +
  geom_histogram(aes(x = weight_kg), binwidth = 100)
```



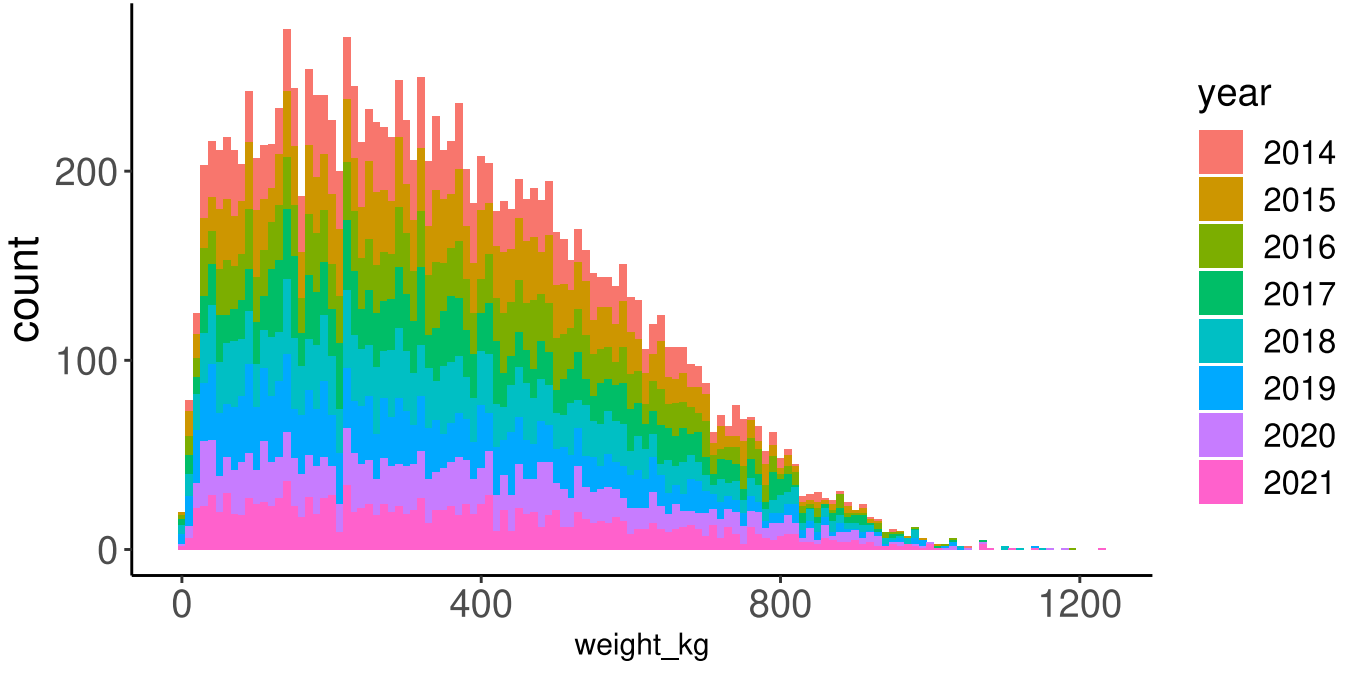
```
pumpS %>%
  filter(type == 'P', est_weight < 9000) %>%
  mutate(weight_kg = (weight_lbs * 10.453592)/1000) %>%
  ggplot() +
  geom_histogram(aes(x = weight_kg)) +
  labs(x = "Weight (tn)")
```



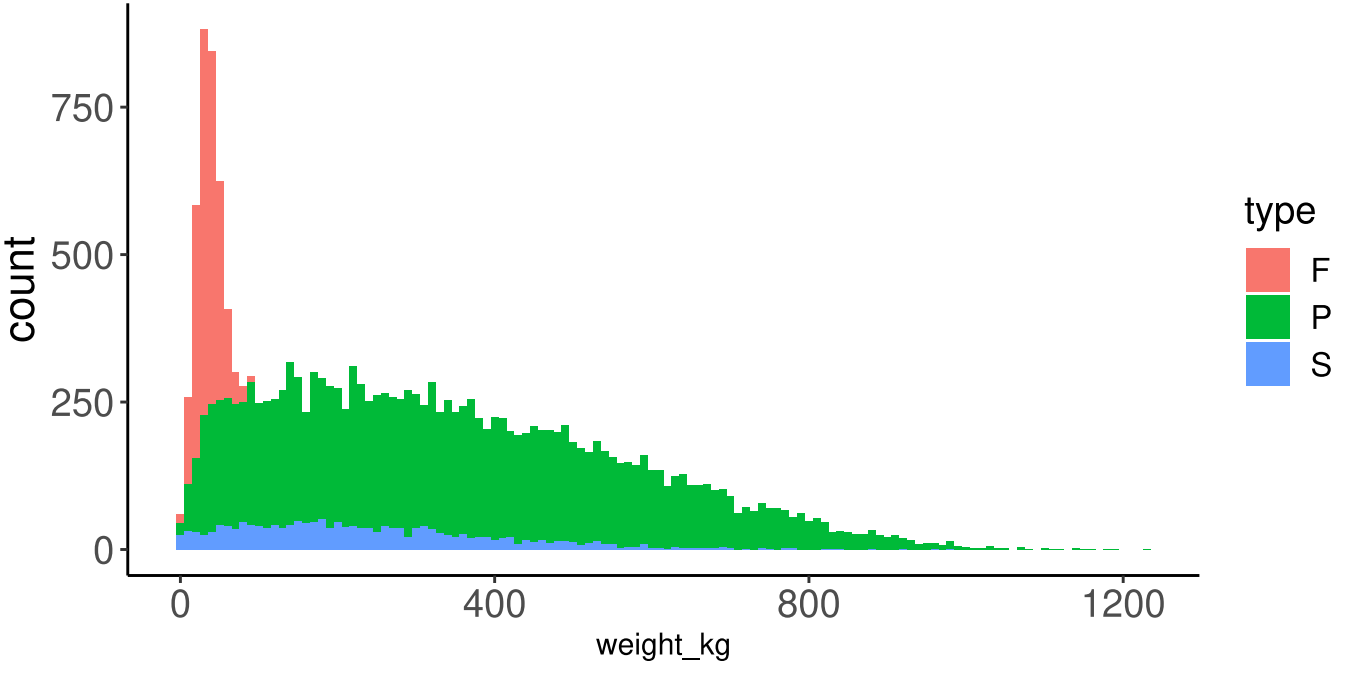
```
pumpS %>%
  filter(type == 'P', est_weight < 9000) %>%
  mutate(weight_kg = weight_lbs * 0.453592) %>%
  ggplot() +
  geom_histogram(aes(x = weight_kg))
```



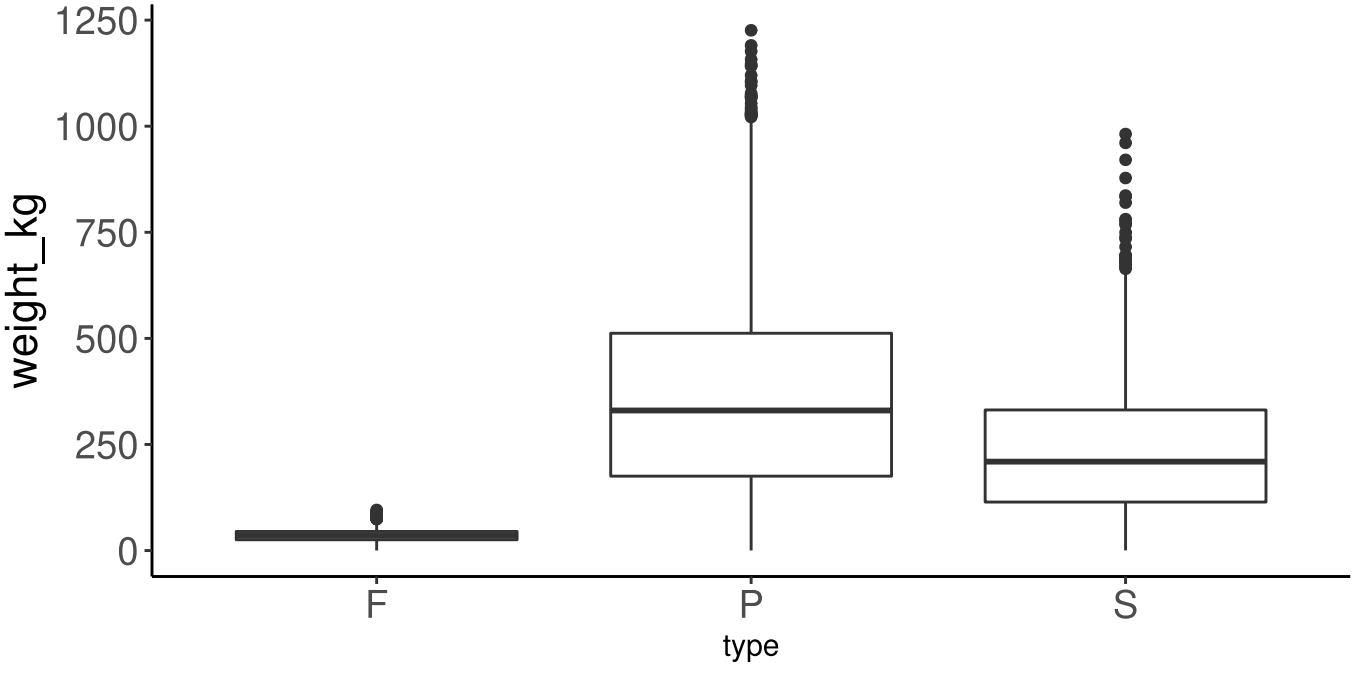
```r
pumpS %>%
  filter(type == 'P', est_weight < 9000) %>%
  mutate(weight_kg = weight_lbs * 0.453592) %>%
  ggplot() +
  geom_histogram(aes(x = weight_kg, fill = year), binwidth = 10)
```



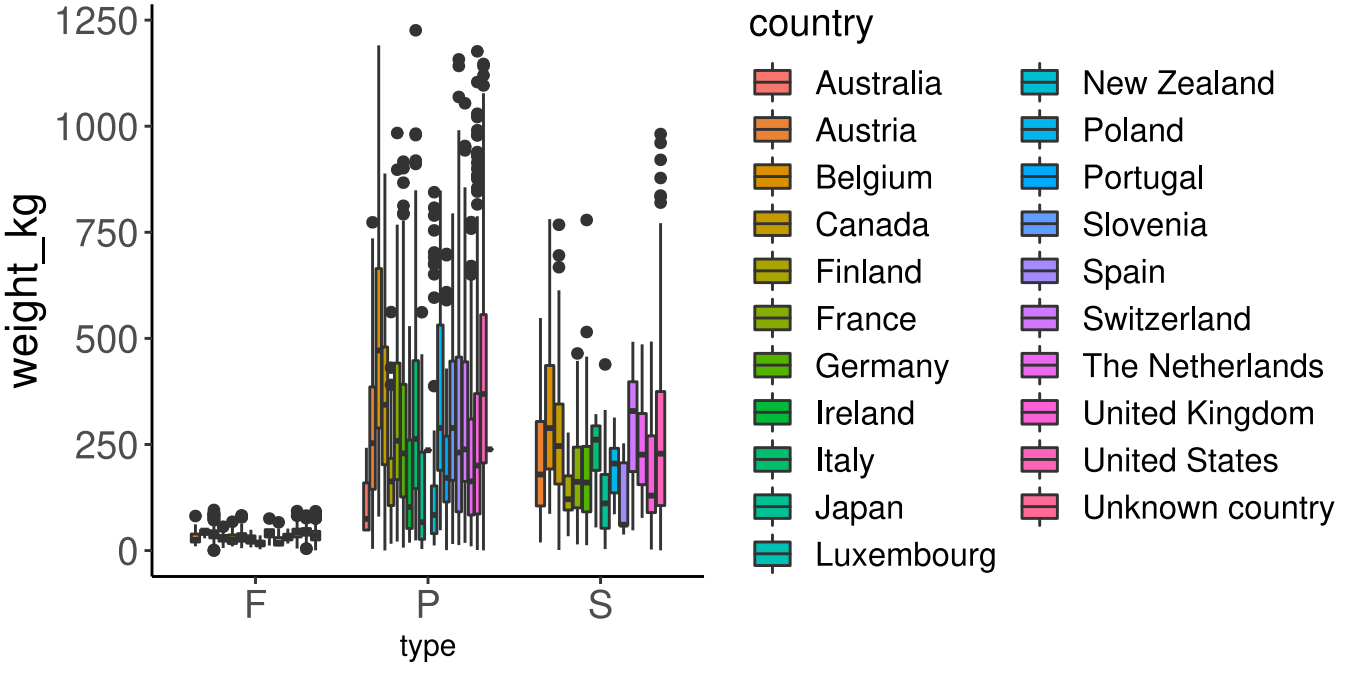
```r
pumpS %>%
  filter(type %in% c('P', 'F', 'S')) %>%
  mutate(weight_kg = weight_lbs * 0.453592) %>%
  ggplot() +
  geom_histogram(aes(x = weight_kg, fill = type), binwidth = 10)
```



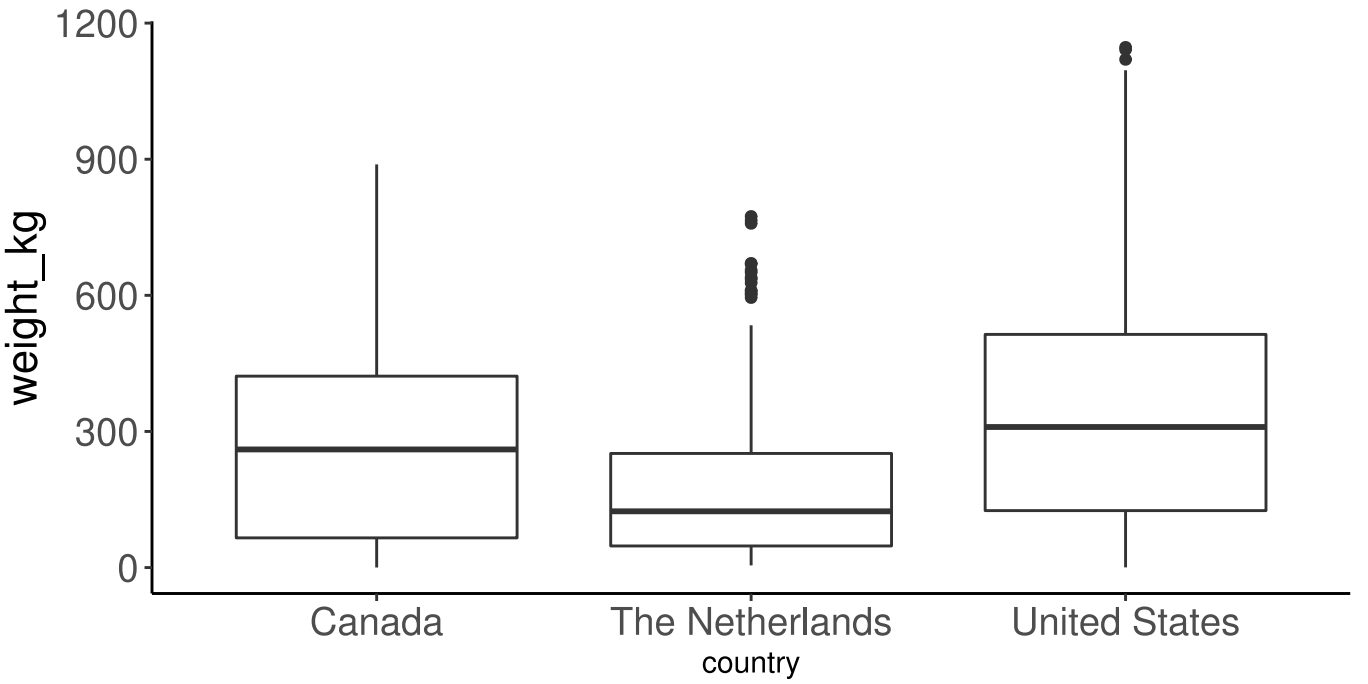
```
pumpS %>%
  filter(type %in% c('P', 'F', 'S')) %>%
  mutate(weight_kg = weight_lbs * 0.453592) %>%
  ggplot() +
  geom_boxplot(aes(x = type, y = weight_kg))
```



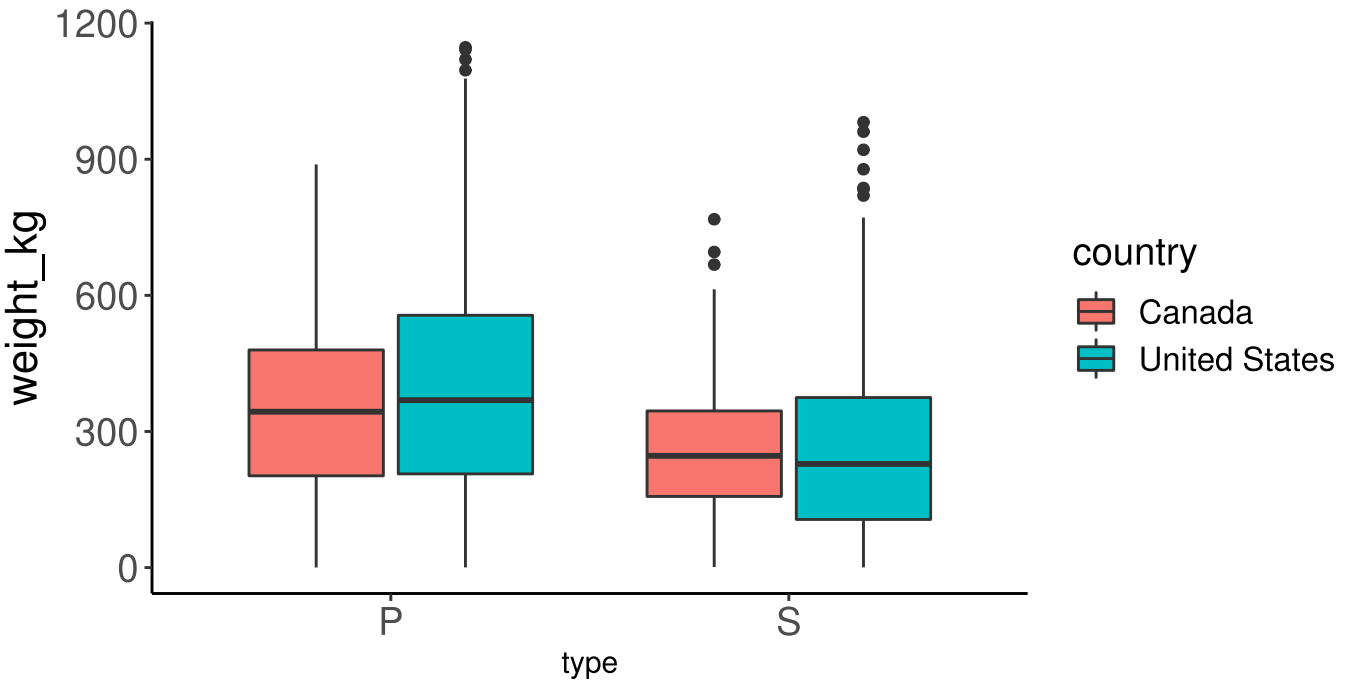
```
pumpS %>%
  filter(type %in% c('P', 'F', 'S')) %>%
  mutate(weight_kg = weight_lbs * 0.453592) %>%
  ggplot() +
  geom_boxplot(aes(x = type, y = weight_kg, fill = country))
```



```
pumpS %>%
  filter(type %in% c('P', 'F', 'S'), country %in% c('Canada','United States', 'The Netherlands')) %>%
  mutate(weight_kg = weight_lbs * 0.453592) %>%
  ggplot() +
  geom_boxplot(aes(x = country, y = weight_kg))
```



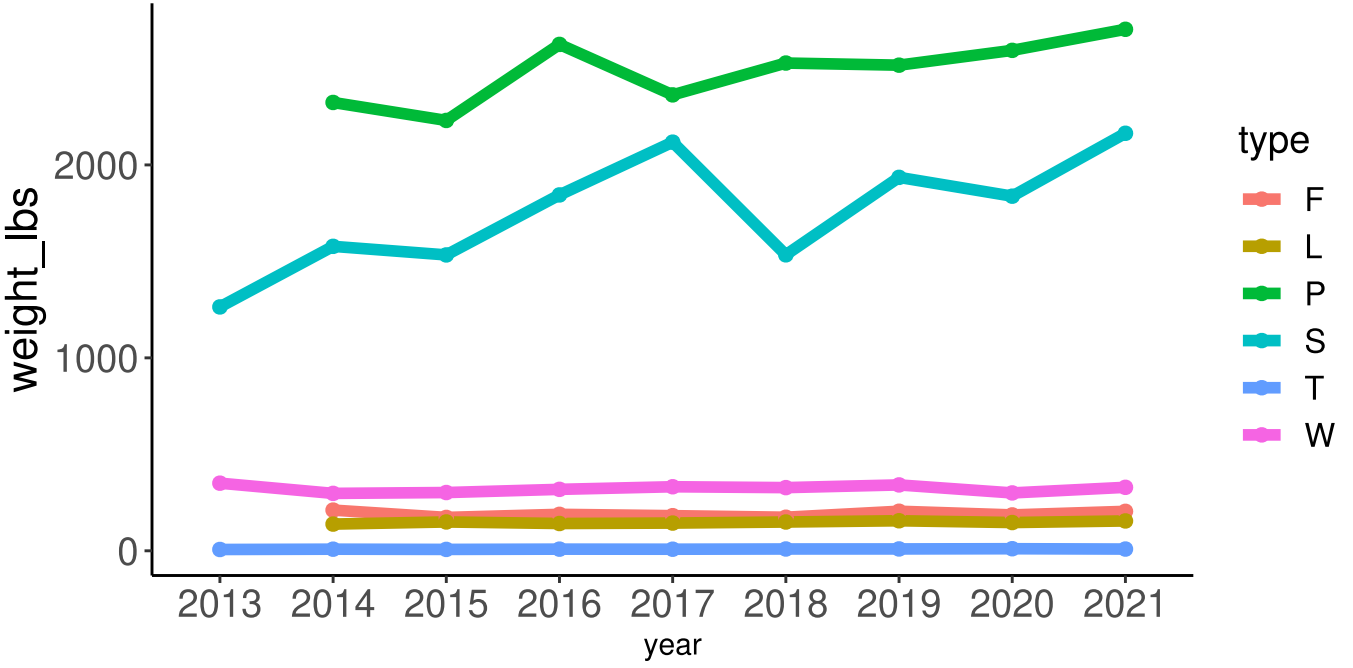
```
pumpS %>%
  filter(type %in% c('P', 'S'), country %in% c('Canada','United States')) %>%
  mutate(weight_kg = weight_lbs * 0.453592) %>%
  ggplot() +
  geom_boxplot(aes(x = type, y = weight_kg, fill = country))
```



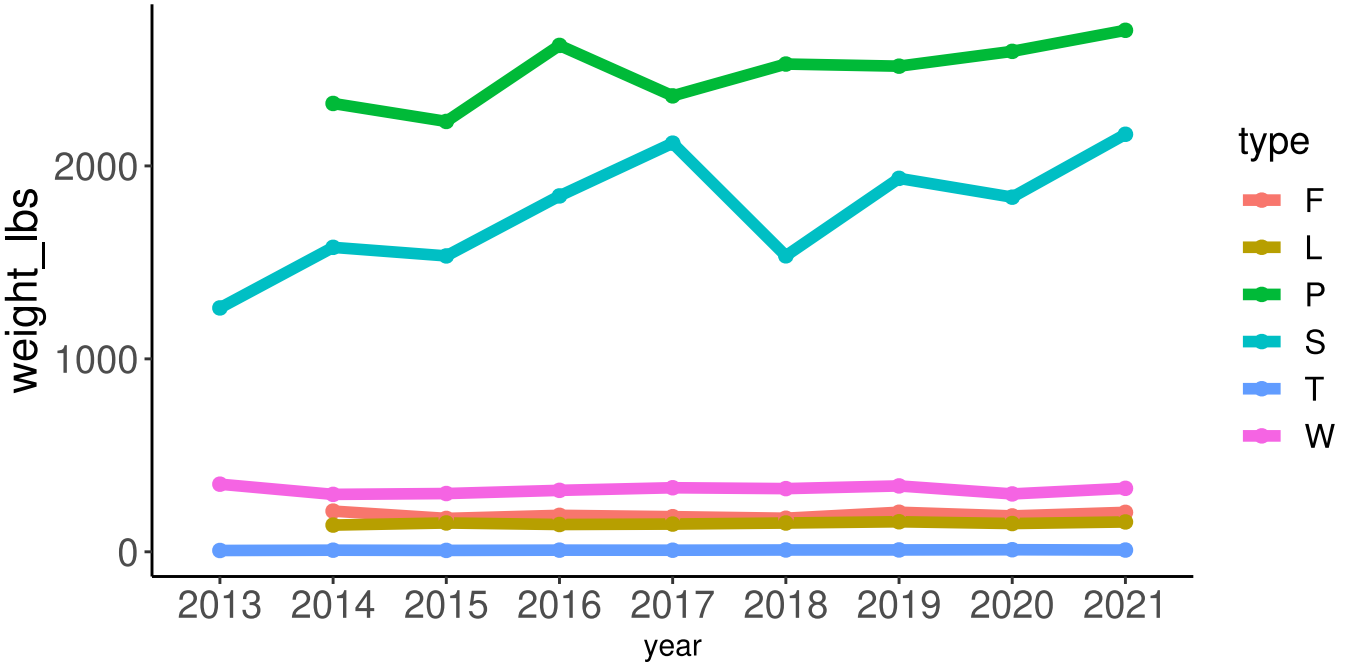
```
pumpS %>%
  filter(place == 1)
```

```
# A tibble: 51 × 14
   year  type  place weight_lbs grower_name    city  state_prov country gpc_site
   <chr> <chr> <chr>      <dbl> <chr>          <chr> <chr>      <chr>   <chr>   
 1 2013  S     1           1264 Pierpont, Edw… Jeff… Maine      United… Damaris…
 2 2013  T     1              6.83 Boudyo, Fabri… Cars… Other      France  Early T…
 3 2013  W     1            350.   Kent, Chris    Seut… Tennessee  United… Operati…
 4 2014  F     1            211    MacKinnon, Jo… Stra… Nova Scot… Canada  The Gre…
 5 2014  L     1            138.   Clementz, Mark Holly Michigan   United… Ohio Va…
 6 2014  P     1           2324.   Meier, Beni    Pfun… Other      Switze… Europam…
 7 2014  S     1           1578    Holub, Scott   Euge… Oregon     United… Baumans…
 8 2014  T     1              8.41 MacCoy, Dan    Ely   Minnesota  United… Early T…
 9 2014  W     1            298.   Gabriele Bart… Nove… Other      Italy   Festa d…
10 2015  F     1            173    Ellenbecker, … Glea… Wisconsin  United… Nekoosa…
# … with 41 more rows, and 5 more variables: seed_mother <chr>,
#   pollinator_father <chr>, ott <dbl>, est_weight <dbl>, pct_chart <dbl>
```

```
pumpS %>%
  filter(place == 1) %>%
  ggplot(aes(x = year, y = weight_lbs, color = type, group = type)) +
  geom_line(size = 2) +
  geom_point(size = 2)
```



```
pumpS %>%
  filter(place == 1, ) %>%
  ggplot(aes(x = year, y = weight_lbs, color = type, group = type)) +
  geom_line(size = 2) +
  geom_point(size = 2)
```



```
pumpYT %>%
  group_by(year) %>%
  mutate(types = paste(type, collapse = '_')) %>%
  select(year, types) %>%
  distinct() -> typesGR
typesGR
```

```
# A tibble: 9 × 2
# Groups:   year [9]
  year  types
  <chr> <chr>
1 2013  S_T_W
2 2014  F_L_P_S_T_W
3 2015  F_L_P_S_T_W
4 2016  F_L_P_S_T_W
5 2017  F_L_P_S_T_W
6 2018  F_L_P_S_T_W
7 2019  F_L_P_S_T_W
8 2020  F_L_P_S_T_W
9 2021  F_L_P_S_T_W
```

```
separate_rows(typesGR, types,sep = '_')
```

```
# A tibble: 51 × 2
# Groups:   year [9]
   year  types
   <chr> <chr>
 1 2013  S
 2 2013  T
 3 2013  W
 4 2014  F
 5 2014  L
 6 2014  P
 7 2014  S
 8 2014  T
 9 2014  W
10 2015  F
# … with 41 more rows
```