Are K-Pop idols debuting too young nowadays?

Flavia Jiang

2023-03-02

Note: Not all graphs created in this file were presented in the article. For some reasons, Chinese characters failed to display in the code chunks when knitted to pdf. However, they still appeared in the graphs.

Default styles for ggplot2

```
style <- theme_light() +
  theme(
    axis.title = element_text(color = "#737373", size = 11),
    axis.text = element_text(size = 9, color = "#737373"),
    legend.title = element_text(color = "#737373", size = 11),
    legend.text = element_text(color = "#737373", size = 9),
    plot.title = element_text(
      color = "#737373",
      size = 15,
      margin = margin(10, 0, 10, 0)
    ),
    axis.title.x = element_text(margin = margin(
      t = 7,
      r = 0,
      b = 7,
      l = 0
    )),
    axis.title.y = element_text(margin = margin(
      t = 0,
      r = 7,
      b = 0,
      l = 7
    )),
    plot.caption = element_text(color = "#737373", size = 9)
  )
```

Data cleaning

Data by idol

```r
dat <- read.csv("debut_age.csv")
# The memberdat dataframe: one idol per row
memberdat <- pivot_longer(dat,
                          cols = 6:28,
                          names_to = "member",
                          values_to = 'age')
memberdat <- filter(memberdat,!is.na(age))
head(memberdat)
```

```
## # A tibble: 6 x 7
##   group             company gender show  debut_year member       age
##   <chr>             <chr>   <chr>  <chr>      <int> <chr>      <int>
## 1 S.E.S.            SM      F      N           1997 debut_age1    17
## 2 S.E.S.            SM      F      N           1997 debut_age2    16
## 3 S.E.S.            SM      F      N           1997 debut_age3    16
## 4 Girls' Generation SM      F      N           2007 debut_age1    18
## 5 Girls' Generation SM      F      N           2007 debut_age2    18
## 6 Girls' Generation SM      F      N           2007 debut_age3    18
```

Data by group

Variables:

- average: average debut age of the group
- numMember: number of members
- ageSum: sum of debut ages of the members
- sd: standard deviation of debut ages
- adult: number of members aged 18 or above when debutting / numMember

```r
# The groupdat dataframe: one group per row
groupdat <-
  summarise(
    group_by(memberdat, group),
    average = mean(age),
    numMember = n(),
    ageSum = sum(age),
    sd = sd(age),
    company = unique(company)[1],
    gender = unique(gender)[1],
    show = unique(show)[1],
    adult = sum(age >= 18) / n(),
    debut_year = unique(debut_year)[1]
  )
head(groupdat)
```

```
## # A tibble: 6 x 10
##   group   average numMember ageSum    sd company gender show  adult debut_year
##   <chr>     <dbl>     <int>  <int> <dbl> <chr>   <chr>  <chr> <dbl>      <int>
## 1 2AM        19.8         4     79 2.22  JYP     M      N      0.75       2008
## 2 2NE1       20.8         4     83 5.06  YG      F      N      0.75       2009
## 3 2PM        19.4         7    136 1.13  JYP     M      N      1          2008
```

```
## 4 4Minute    17.8         5     89 1.79  CUBE    F      N      0.6         2009
## 5 AOA        19.1         8    153 1.96  OTHER   F      N      0.75        2012
## 6 ATEEZ      19.1         8    153 0.641 OTHER   M      N      1           2018
```

Basic statistics (overall)

Distribution of debut age

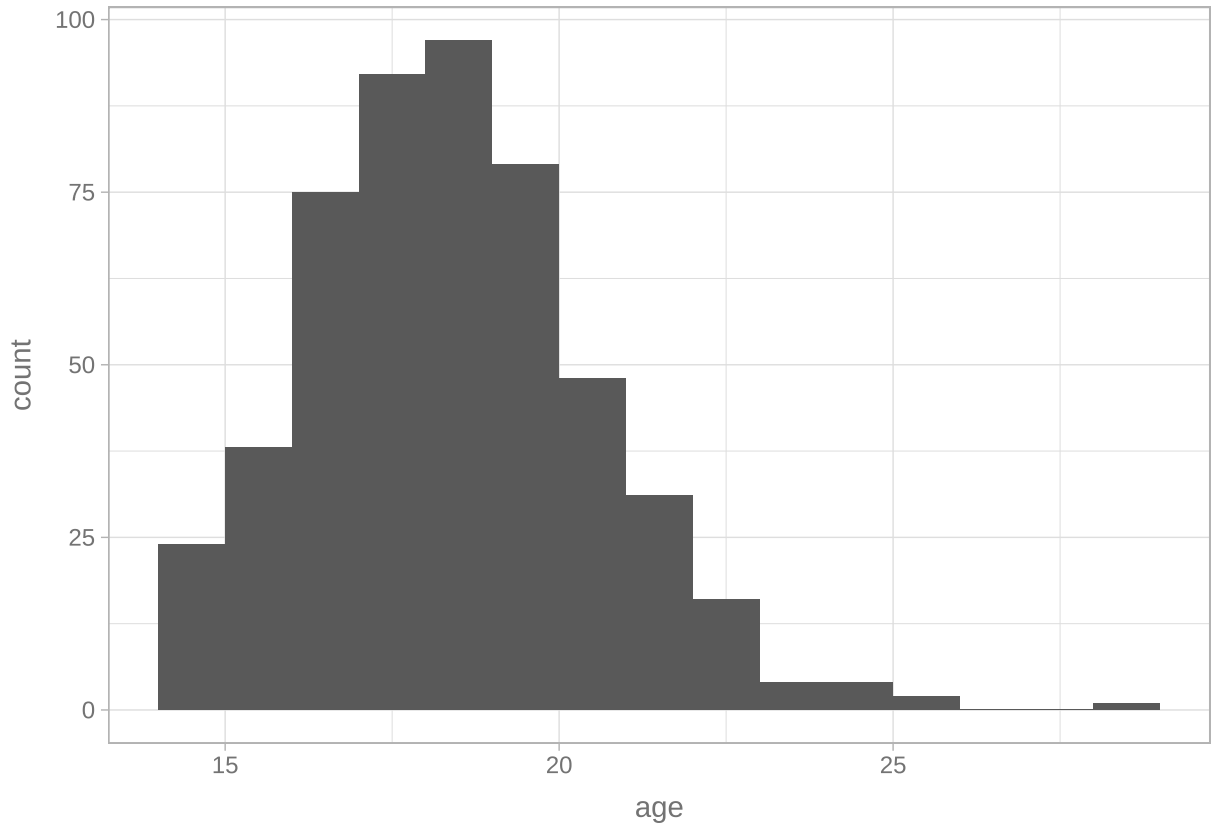
```
median(memberdat$age)
```

```
## [1] 19
```

```
mean(memberdat$age)
```

```
## [1] 18.89041
```

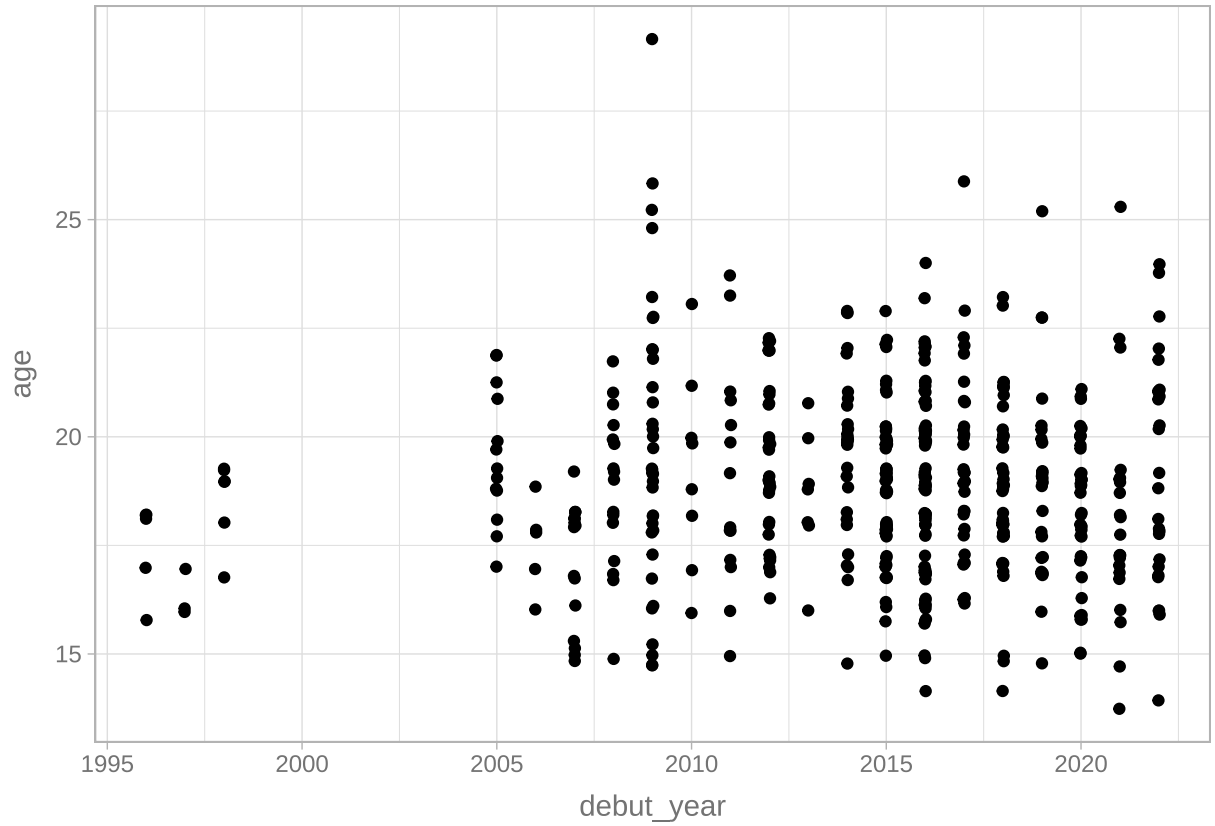
```
ggplot(data = memberdat, aes(x = age)) + geom_histogram(breaks = 14:29) + style
```



Distribution of debut age by debut year

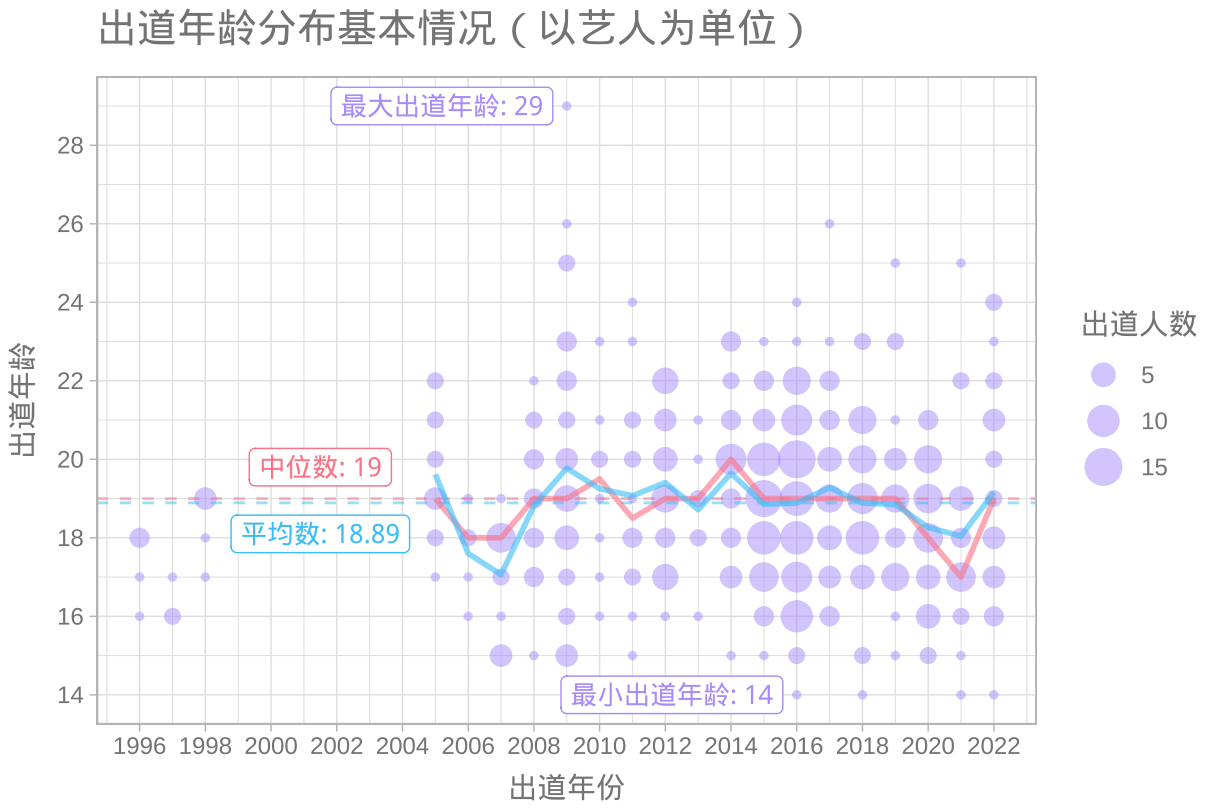
```r
year_mean <-
  summarize(group_by(memberdat, debut_year),
            average = mean(age),
            count = n())[4:21, ]
year_median <-
  summarize(group_by(memberdat, debut_year),
            median = median(age),
            count = n())[4:21, ]

ggplot(data = memberdat, aes(x = debut_year, y = age)) +
  geom_jitter(width = 0.02, height = 0.3) +
  style
```



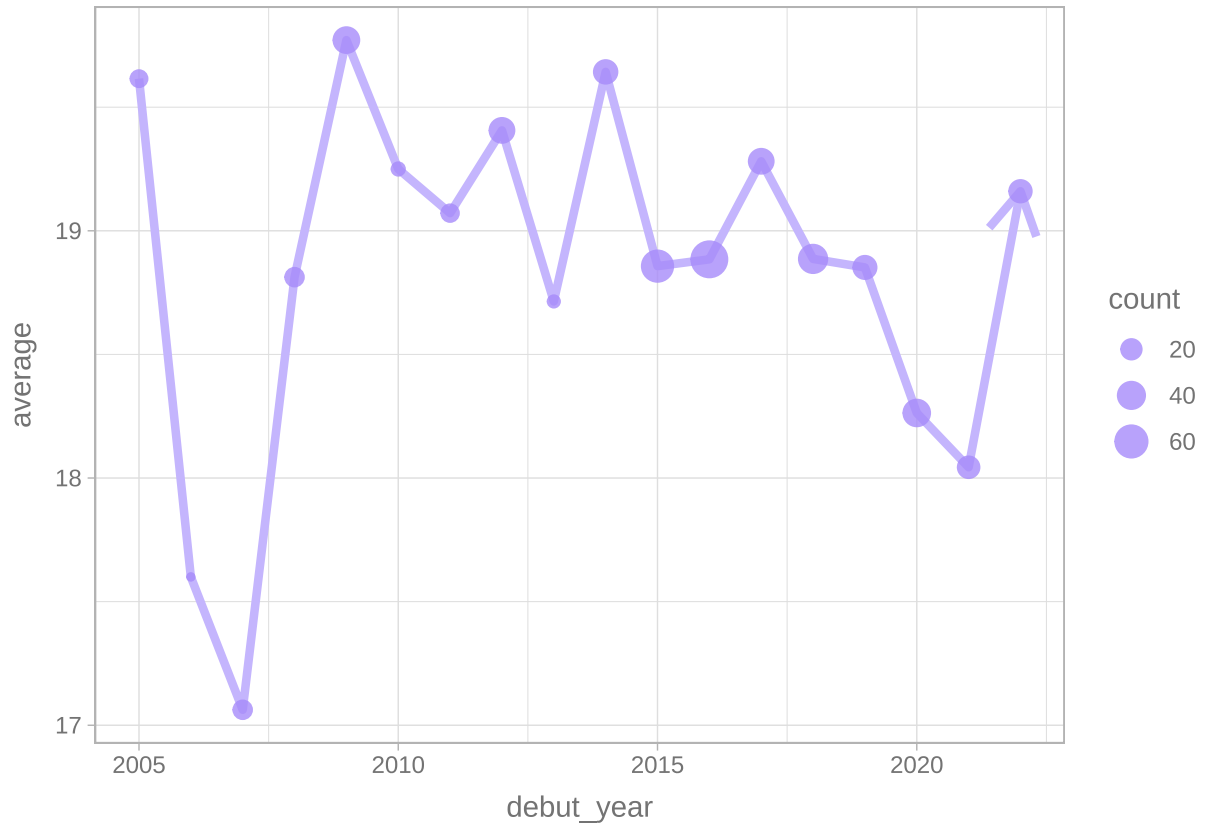

```r
sum_member <-
  summarize(group_by(memberdat, debut_year, age), count = n())
ggplot() +
  geom_point(
    data = sum_member,
    aes(x = debut_year, y = age, size = count),
    color = "#a78bfa",
    alpha = 0.5
  ) +
  style +
  geom_hline(
    yintercept = c(19, 18.89),
```

```
  linetype = "dashed",
  size  = 0.5,
  color = c("#fb7185", "#38bdf8"),
  alpha = 0.5
) +
geom_path(
  data = year_median,
  color = "#fb7185",
  alpha = 0.6,
  size = 1,
  aes(x = debut_year, y = median)
)  + labs(x = "  ",
          y = "  ",
          size = "  ",
          title = "          ") +
geom_path(
  data = year_mean,
  color = "#38bdf8",
  alpha = 0.6,
  size = 1,
  aes(x = debut_year, y = average)
) +
geom_label(aes(x = 2001.5, y = 19.8, label = "  : 19"),
           color = "#fb7185",
           size = 3.5) +
geom_label(
  aes(x = 2001.5, y = 18.1, label = "  : 18.89"),
  color = "#38bdf8",
  size = 3.5
) +
geom_label(aes(x  = 2005.2, y = 29, label = "    : 29"),
           color = "#a78bfa",
           size = 3.5) +
geom_label(aes(x = 2012.2, y = 14, label = "    : 14"),
           color = "#a78bfa",
           size = 3.5) +
scale_y_continuous(breaks = seq(from = 14, to = 28, by = 2)) +
scale_x_continuous(breaks = seq(from = 1996, to = 2022, by = 2))
```

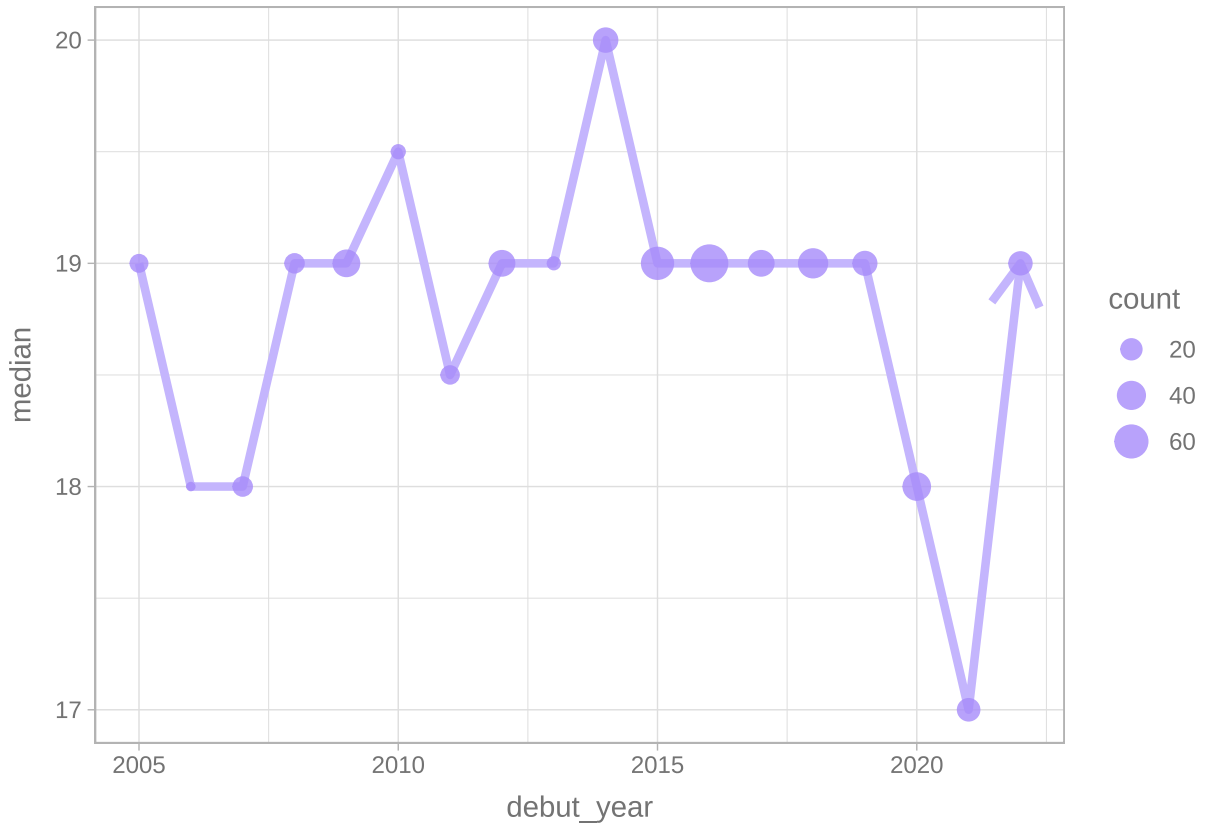



```
ggsave("debut_age_distribution.png",
       plot = last_plot(),
       device = png)

ggplot(data = year_mean) +
  geom_path(
    color = "#c4b5fd",
    arrow = arrow(),
    size = 1.5,
    aes(x = debut_year, y = average)
  ) +
  geom_point(
    aes(x = debut_year, y = average, size = count),
    color = "#a78bfa",
    alpha = .8
  )  +
  style
```

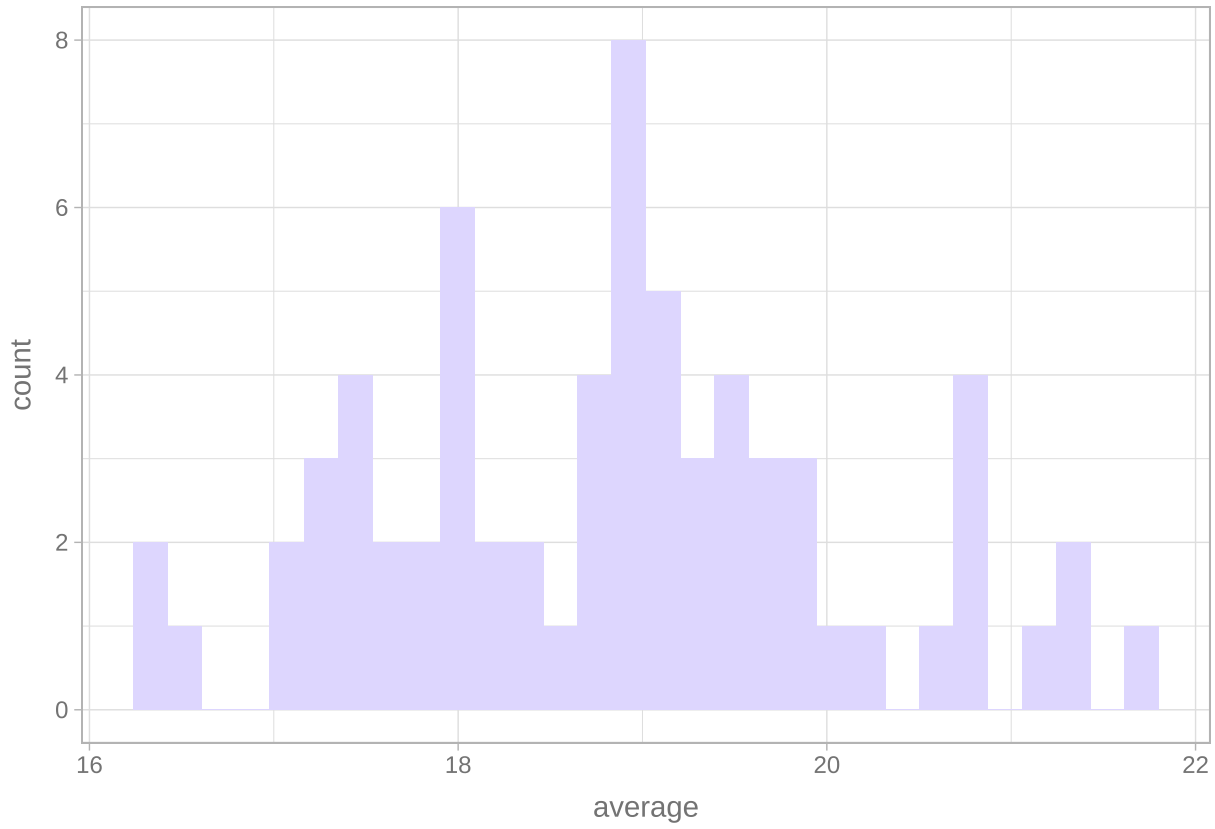



```
ggplot(data = year_median) +
  geom_path(
    color = "#c4b5fd",
    arrow = arrow(),
    size = 1.5,
    aes(x = debut_year, y = median)
  ) +
  geom_point(
    aes(x = debut_year, y = median, size = count),
    color = "#a78bfa",
    alpha = .8
  )  +
  style
```



Distribution of group average debut ages

```
ggplot(data = groupdat, aes(x = average)) +
  geom_histogram(fill = "#ddd6fe") +
  style
```

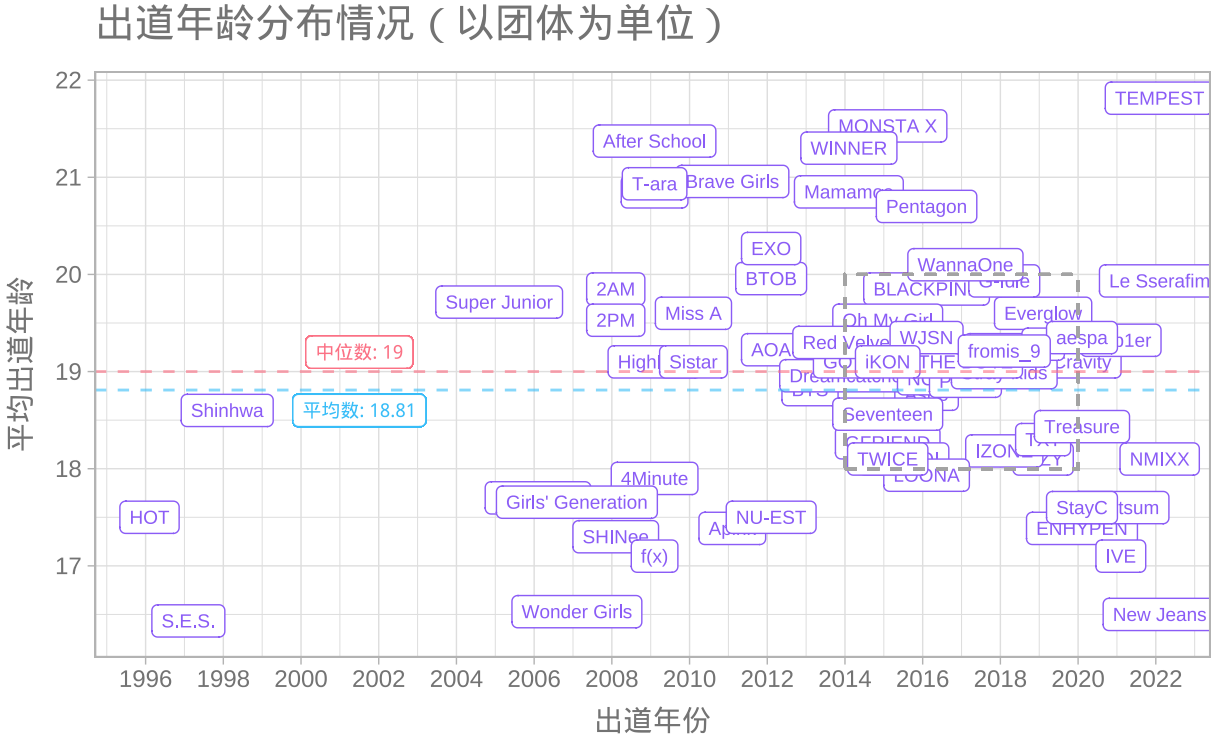



```
head(groupdat[order(groupdat$average), ])
```

```
## # A tibble: 6 x 10
##   group     average numMember ageSum    sd company gender show  adult debut_year
##   <chr>       <dbl>     <int>  <int> <dbl> <chr>   <chr>  <chr> <dbl>       <int>
## 1 S.E.S.       16.3         3     49 0.577 SM      F      N     0            1997
## 2 New Jeans    16.4         5     82 1.52  HYBE    F      N     0.2          2022
## 3 Wonder G~    16.4         7    115 1.81  JYP     F      N     0.429        2007
## 4 IVE          17           6    102 1.67  STARSH~ F      N     0.333        2021
## 5 f(x)         17           5     85 2.92  SM      F      N     0.2          2009
## 6 SHINee       17.2         5     86 1.48  SM      M      N     0.4          2008
```

```
ggplot(data = groupdat, aes(x = debut_year, y = average)) +
  geom_point(color = "#8b5cf6") +
  geom_label(
    label = groupdat$group,
    nudge_x = 0.1,
    nudge_y = 0.1,
    size = 2.5,
    color = "#8b5cf6"
  ) +
  labs(
    x = "  ",
    y = "   ",
    title = "         ",
```

```
  caption = " : NCT  NCT DREAM    15.6      \n :        "
) +
style +
scale_y_continuous(breaks = 16:22) +
scale_x_continuous(breaks = seq(from = 1996, to = 2022, by = 2)
) +
geom_rect(
  aes(
    xmin = 2014,
    xmax = 2020,
    ymin = 18,
    ymax = 20
  ),
  fill = NA,
  color = "#a3a3a3",
  linetype = "dashed"
) +
geom_label(aes(x = 2001.5, y = 19.2, label = "  : 19"),
           color = "#fb7185",
           size = 2.5) +
geom_label(
  aes(x = 2001.5, y = 18.6, label = "  : 18.81"),
  color = "#38bdf8",
  size = 2.5
) +
geom_hline(
  yintercept = c(19, 18.81),
  linetype = "dashed",
  size = 0.5,
  color = c("#fb7185", "#38bdf8"),
  alpha = 0.6
)
```

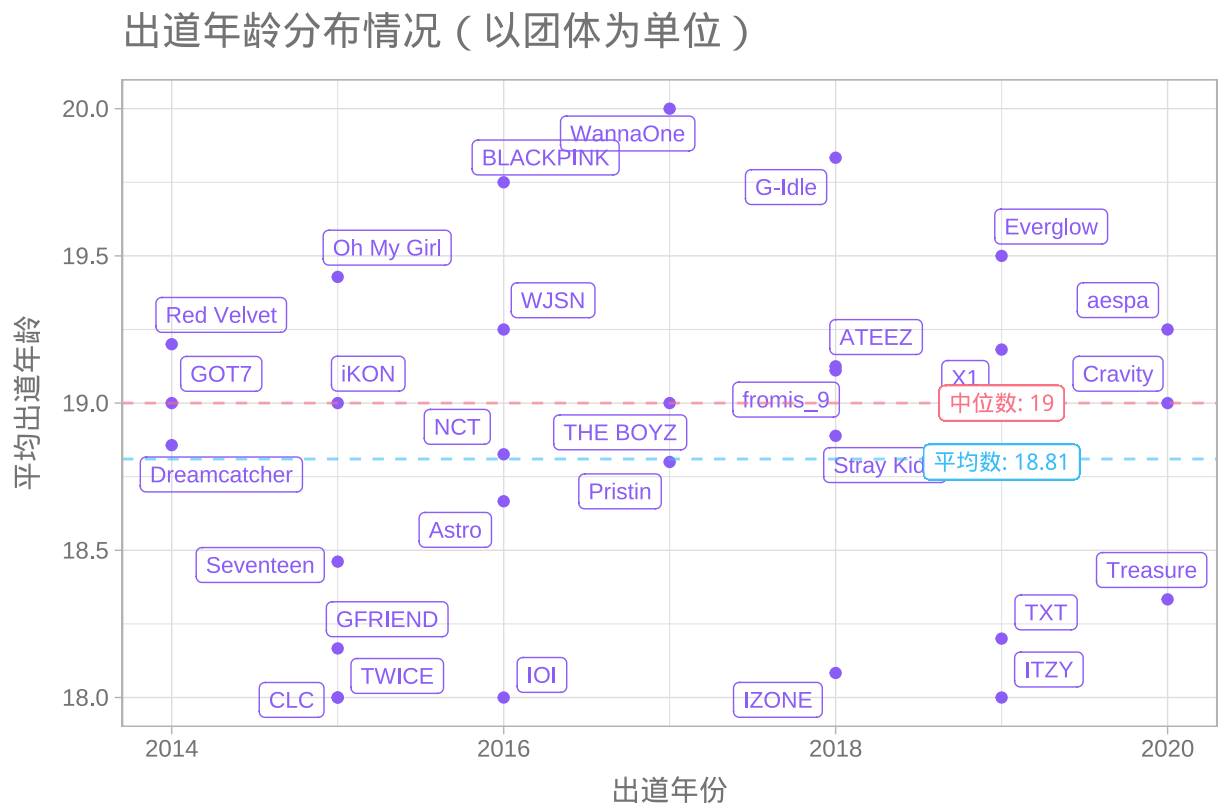


注: NCT的分队NCT DREAM平均出道年龄为15.6。此处只显示大队数据。
注: 长方形部分可见下方放大图。

```
ggsave("group_mean_debut_age.png",
       plot = last_plot(),
       device = png)

ggplot(data = groupdat, aes(x = debut_year, y = average)) +
  geom_point(color = "#8b5cf6") +
  geom_label_repel(
    label = groupdat$group,
    size = 3,
    color = "#8b5cf6",
    fill = NA,
    max.overlaps = 20,
    min.segment.length = 1
  ) +
  ylim(c(18, 20)) + xlim(c(2014, 2020)) +
  style +
  labs(x = "  ", y = "    ", title = "         ") +
  geom_hline(
    yintercept = c(19, 18.81),
    linetype = "dashed",
    size = 0.5,
    color = c("#fb7185", "#38bdf8"),
    alpha = 0.6
  ) +
  geom_label(aes(x = 2019, y = 19, label = "  : 19"),
             color = "#fb7185",
```

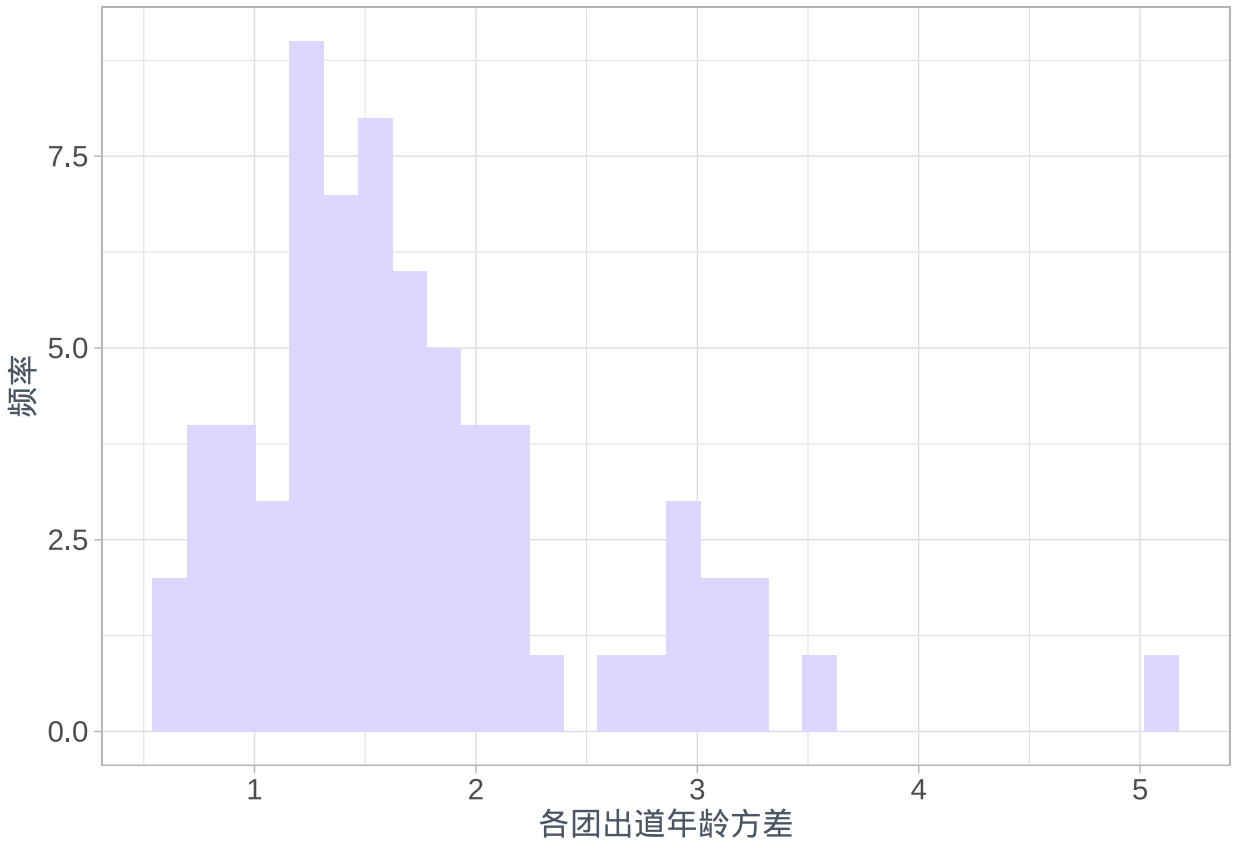
```
            size = 3) +
  geom_label(aes(x = 2019, y = 18.8, label = "  : 18.81"),
             color = "#38bdf8",
             size = 3)
```




```
ggsave("group_mean_debut_age_zoomed_in.png",
       plot = last_plot(),
       device = png)
```

Distribution of debut age standard deviations within groups

```
ggplot(data = groupdat, aes(x = sd)) +
  geom_histogram(fill = "#ddd6fe") +
  labs(x = "    ",  y = " ") +
  theme_light() +
  theme(axis.title = element_text(colour = "#4b5563", size = 12),
        axis.text = element_text(size = 11))
```




```
filter(groupdat, sd > 3) # 2NE1, After School, Le Sserafim, WannaOne, X1, Kepler
```

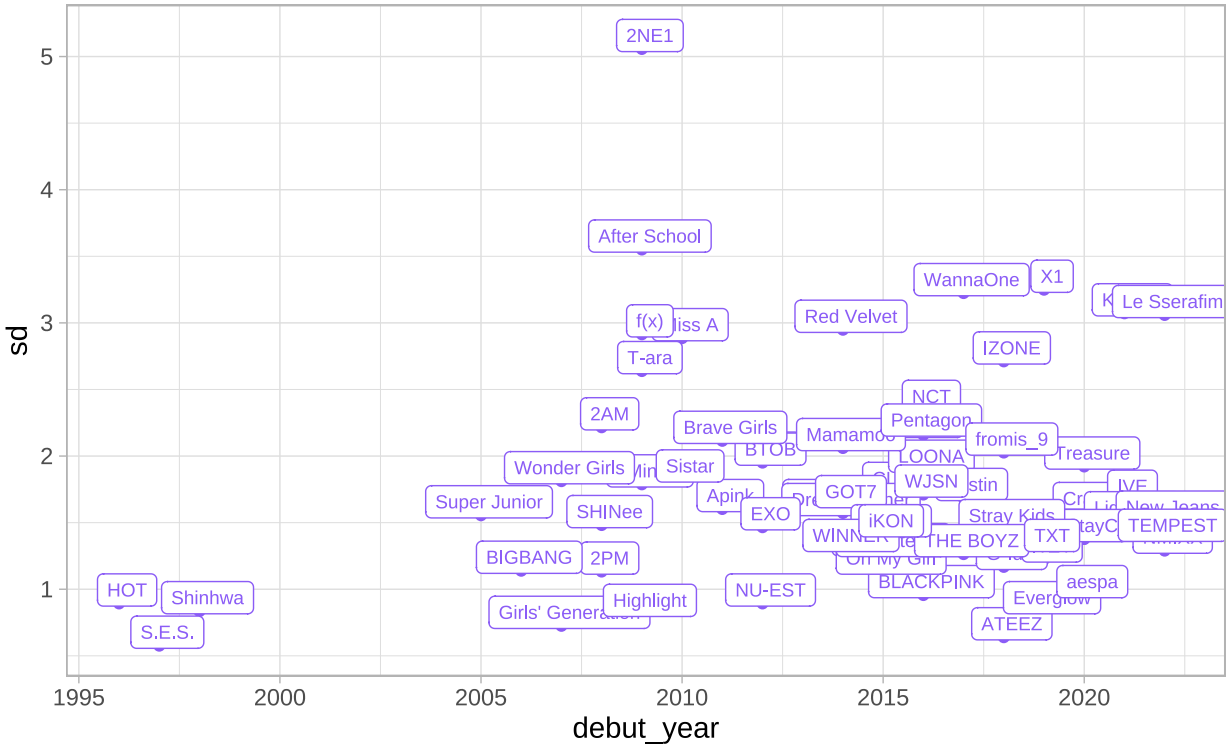
```
## # A tibble: 6 x 10
##   group     average numMember ageSum    sd company gender show  adult debut_year
##   <chr>       <dbl>     <int>  <int> <dbl> <chr>   <chr>  <chr> <dbl>       <int>
## 1 2NE1         20.8         4     83  5.06 YG      F      N     0.75         2009
## 2 After Sc~    21.3        11    234  3.55 HYBE    F      N     1            2009
## 3 Kep1er       19.2         9    173  3.07 MNET    F      Y     0.556        2021
## 4 Le Ssera~    19.8         6    119  3.06 HYBE    F      N     0.667        2022
## 5 WannaOne     20          11    220  3.22 MNET    M      Y     0.727        2017
## 6 X1           19.2        11    211  3.25 MNET    M      Y     0.455        2019
```

```
filter(groupdat, sd < 0.7)
```

```
## # A tibble: 2 x 10
##   group  average numMember ageSum    sd company gender show  adult debut_year
##   <chr>    <dbl>     <int>  <int> <dbl> <chr>   <chr>  <chr> <dbl>       <int>
## 1 ATEEZ     19.1         8    153 0.641 OTHER   M      N         1        2018
## 2 S.E.S.    16.3         3     49 0.577 SM      F      N         0        1997
```

```
ggplot(data = groupdat, aes(x = debut_year, y = sd)) +
  geom_point(color = "#8b5cf6") +
  geom_label(
    label = groupdat$group,
```

```
    nudge_x = 0.2,
    nudge_y = 0.1,
    size = 2.5,
    color = "#8b5cf6",
  ) + theme_light()
```

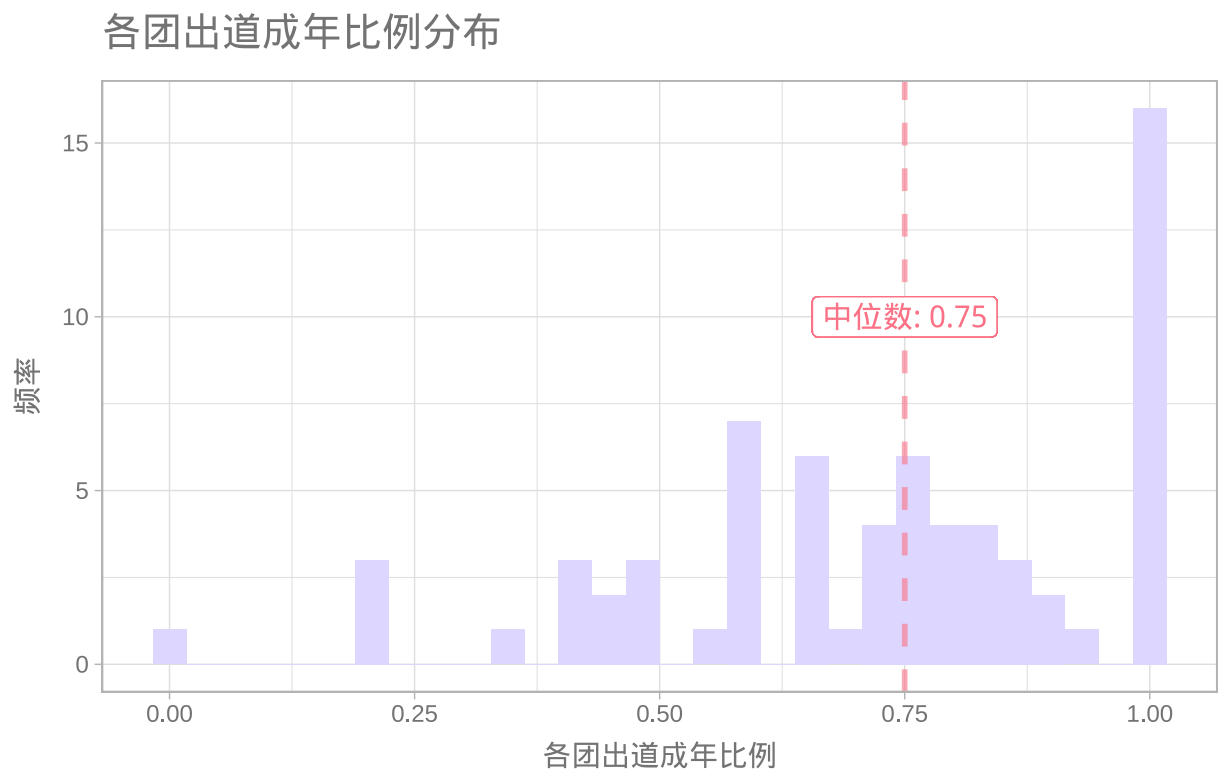



```
summarize(group_by(memberdat, group), diff = max(age) - min(age))
```

```
## # A tibble: 68 x 2
##    group         diff
##    <chr>        <int>
##  1 2AM              5
##  2 2NE1            10
##  3 2PM              3
##  4 4Minute          4
##  5 AOA              6
##  6 ATEEZ            2
##  7 After School    11
##  8 Apink            5
##  9 Astro            6
## 10 BIGBANG          3
## # i 58 more rows
```

Distribution of adult rates

```
ggplot(data = groupdat, aes(x = adult)) +
  geom_histogram( fill = "#ddd6fe") +
  labs( x = "     ",  y = " ", caption = " 18   ", title = "      ") +
  theme_light() +
  style +
  geom_vline(xintercept = 0.75, linetype = "dashed", size = 1, color = c("#fb7185"), alpha = 0.6)+
  geom_label(aes(x = 0.75, y = 10, label = "  : 0.75"), color= "#fb7185", size  = 4)
```



注：18岁为成年。

```
ggsave(
  "adult_percentage.png",
  plot = last_plot(),
  device = png)

filter( groupdat, adult == 1)
```

```
## # A tibble: 16 x 10
##    group    average numMember ageSum    sd company gender show  adult debut_year
##    <chr>      <dbl>     <int>  <int> <dbl> <chr>   <chr>  <chr> <dbl>       <int>
##  1 2PM         19.4         7    136 1.13  JYP     M      N         1        2008
##  2 ATEEZ       19.1         8    153 0.641 OTHER   M      N         1        2018
##  3 After S~    21.3        11    234 3.55  HYBE    F      N         1        2009
##  4 BLACKPI~    19.8         4     79 0.957 YG      F      N         1        2016
##  5 Brave G~    20.9         7    146 2.12  OTHER   F      N         1        2011
##  6 EXO         20.2        12    242 1.47  SM      M      N         1        2012
```

```
##  7 Everglow    19.5         6    117 0.837 OTHER   F      N         1       2019
##  8 G-Idle      19.8         6    119 1.17  CUBE    F      N         1       2018
##  9 Highlig~    19           4     76 0.816 OTHER   M      N         1       2009
## 10 MONSTA X    21.4         7    150 1.27  STARSH~ M      N         1       2015
## 11 Mamamoo     20.8         4     83 2.06  OTHER   F      N         1       2014
## 12 Oh My G~    19.4         7    136 1.13  OTHER   F      N         1       2015
## 13 Pentagon    20.6        10    206 2.17  CUBE    M      N         1       2016
## 14 TEMPEST     21.7         7    152 1.38  OTHER   M      N         1       2022
## 15 WINNER      21.2         5    106 1.30  YG      M      N         1       2014
## 16 aespa       19.2         4     77 0.957 SM      F      N         1       2020
```

```
filter( groupdat, adult == 0) # Only one: S.E.S.
```

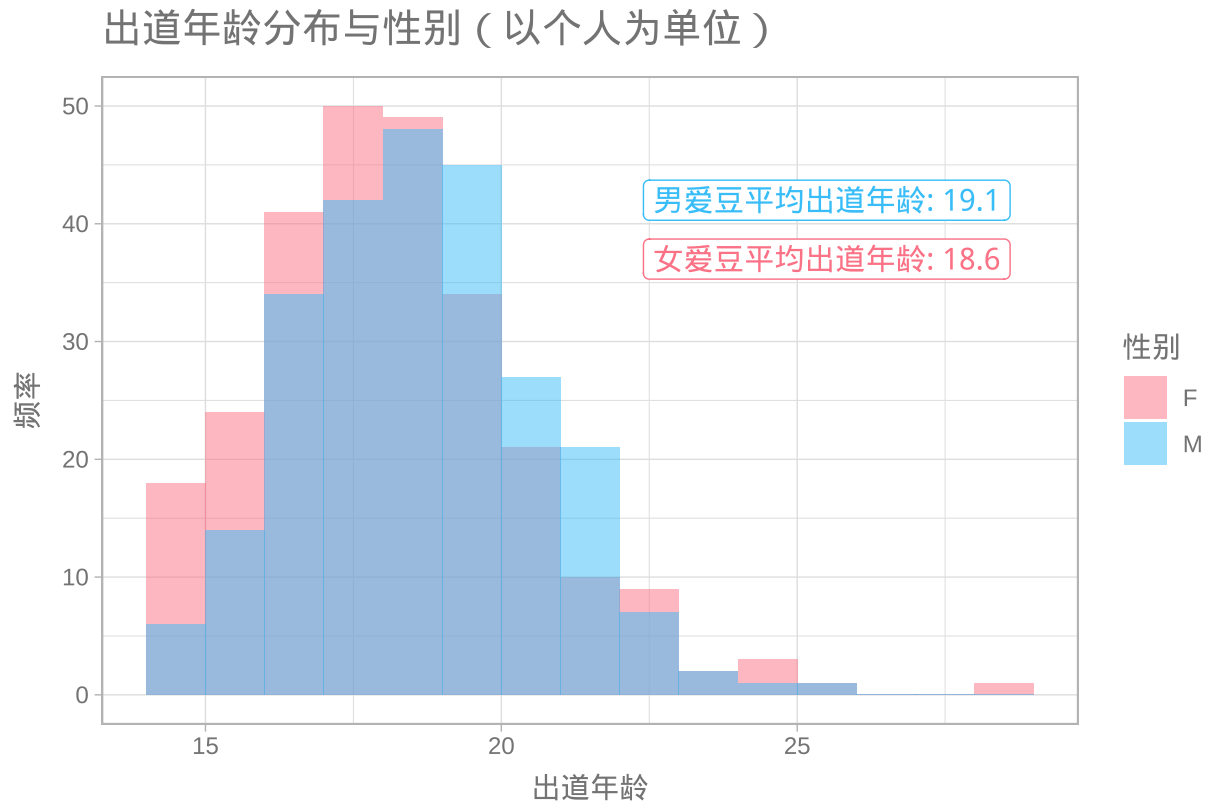
```
## # A tibble: 1 x 10
##   group  average numMember ageSum    sd company gender show  adult debut_year
##   <chr>    <dbl>     <int>  <int> <dbl> <chr>   <chr>  <chr> <dbl>      <int>
## 1 S.E.S.    16.3         3     49 0.577 SM      F      N         0       1997
```

```
median( groupdat$adult)
```

```
## [1] 0.75
```

Analysis by gender

```
ggplot() +
  geom_histogram(
    data = memberdat,
    aes(x = age, fill = gender),
    position = "identity",
    alpha = 0.5,
    breaks = 14:29
  ) +
  style +
  labs(x = "  ",
       y = " ",
       fill = " ",
       title = "          ") +
  geom_label(
    aes(x = 25.5, y  = 42, label = "     : 19.1"),
    color = "#38bdf8",
    size  = 4
  ) +
  geom_label(
    aes(x = 25.5, y = 37, label = "     : 18.6"),
    color = "#fb7185",
    size  = 4
  ) +
  scale_fill_manual(values = c("#fb7185", "#38bdf8"))
```




```
ggsave("debut_age_by_gender_bar.png",
       plot = last_plot(),
       device = png)

mean(filter(memberdat, gender == "F")$age)
```

```
## [1] 18.65019
```

```
median(filter(memberdat, gender == "F")$age)
```

```
## [1] 18
```

```
mean(filter(memberdat, gender == "M")$age)
```

```
## [1] 19.14516
```

```
median(filter(memberdat, gender == "M")$age)
```

```
## [1] 19
```

```
t.test(
  filter(memberdat, gender == "F")$age,
  filter(memberdat, gender == "M")$age,
  alternative = "less"
)
```

```
##
##  Welch Two Sample t-test
##
## data:  filter(memberdat, gender == "F")$age and filter(memberdat, gender == "M")$age
## t = -2.6046, df = 506.3, p-value = 0.004735
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##        -Inf -0.1818089
## sample estimates:
## mean of x mean of y
##  18.65019  19.14516
```

Average debut age

```
dat_f <- filter(groupdat, gender == "F")
sum(dat_f$ageSum) / sum(dat_f$numMember) # 18.6
```

```
## [1] 18.65019
```

```
median(dat_f$average) # 19
```

```
## [1] 18.82857
```

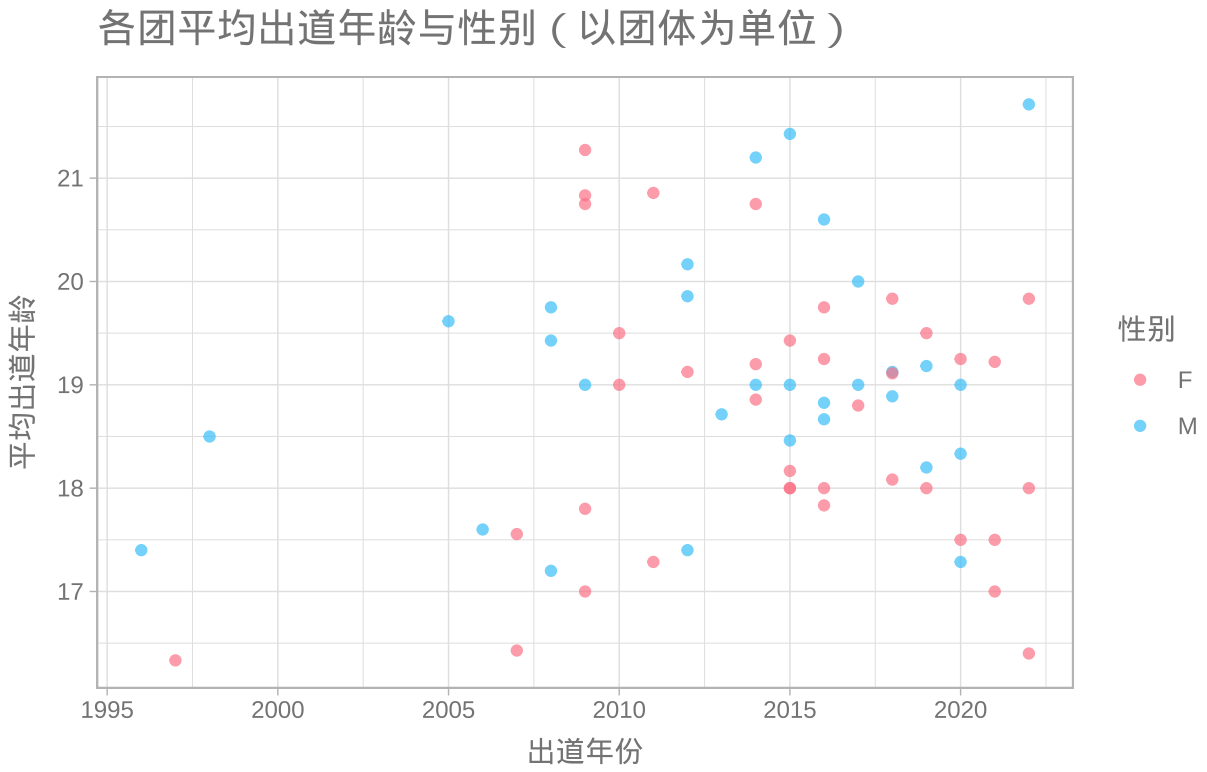
```
dat_m <- filter(groupdat, gender == "M")
sum(dat_m$ageSum) / sum(dat_m$numMember) # 19
```

```
## [1] 19.14516
```

```
median(dat_m$average) # 19
```

```
## [1] 19
```

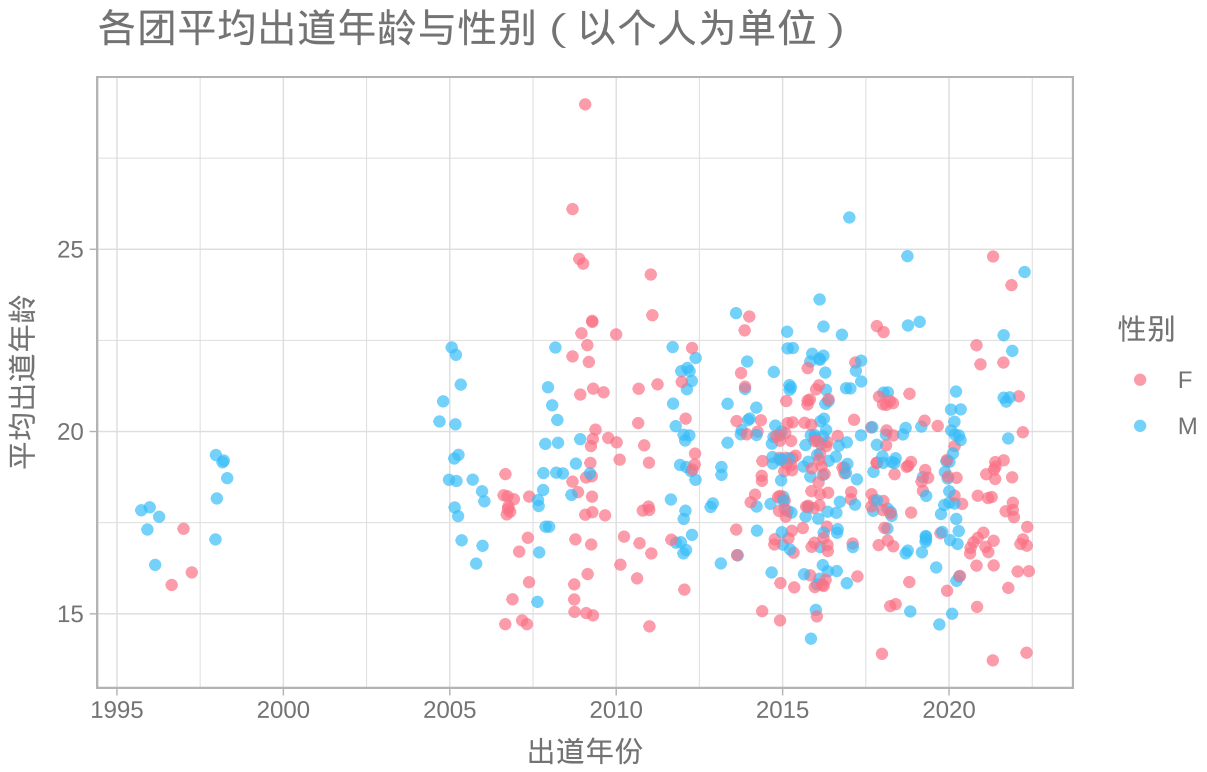
```
ggplot(data = groupdat, aes(x = debut_year, y = average, color = gender)) +
  geom_point(alpha = 0.7) +
  style +
  labs(
    caption = " : NCT  NCT DREAM    15.6    NCT    ",
    title = "          ",
    x = "  ",
    y = "   ",
    color = " "
  ) +
  scale_color_manual(values = c("#fb7185", "#38bdf8"))
```



注: NCT的小分队NCT DREAM出道平均年龄为15.6岁。图上只显示NCT全体的平均年龄。

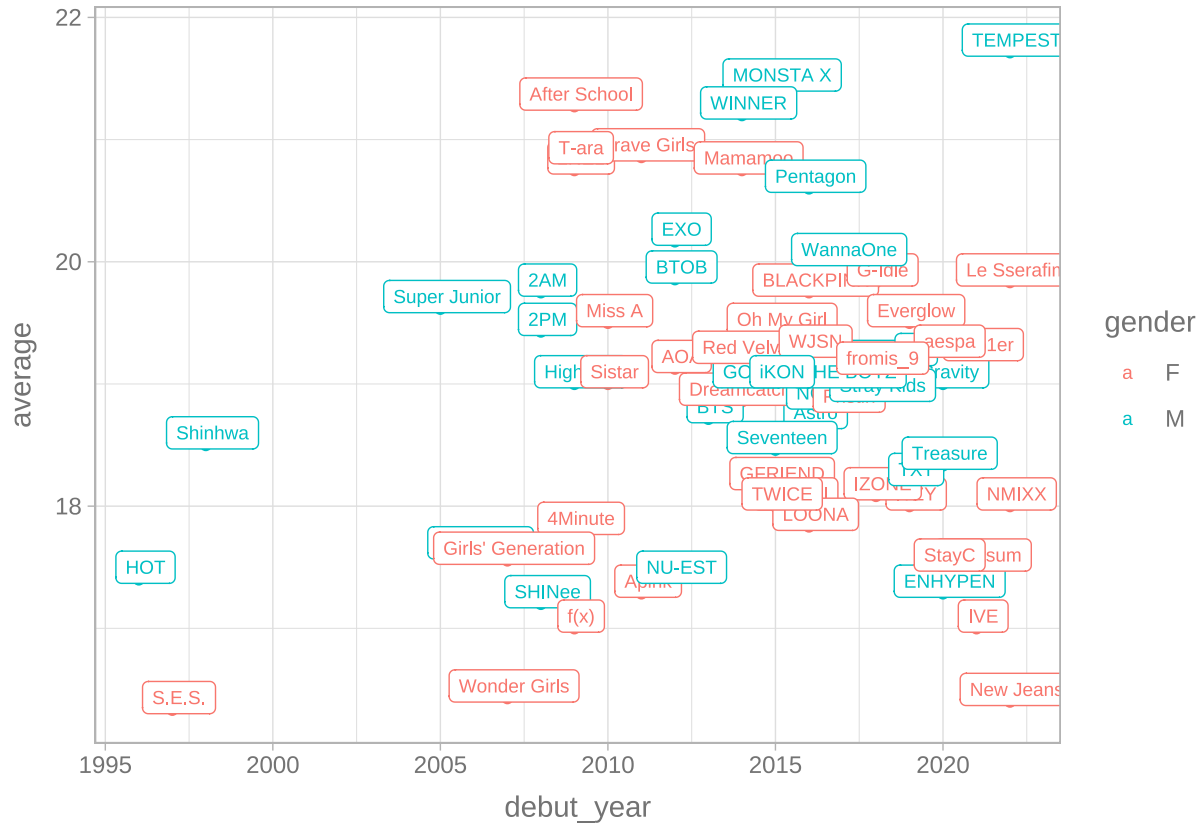
```
ggsave("debut_age_by_gender_scatterplot.png",
       plot = last_plot(),
       device = png)

ggplot(data = memberdat, aes(x = debut_year, y = age, color = gender)) +
  geom_jitter(alpha = 0.7) +
  style +
  labs(
    caption = " : NCT  NCT DREAM    15.6    NCT    ",
    title = "          ",
    x = "  ",
    y = "   ",
    color = " "
  ) +
  scale_color_manual(values = c("#fb7185", "#38bdf8"))
```

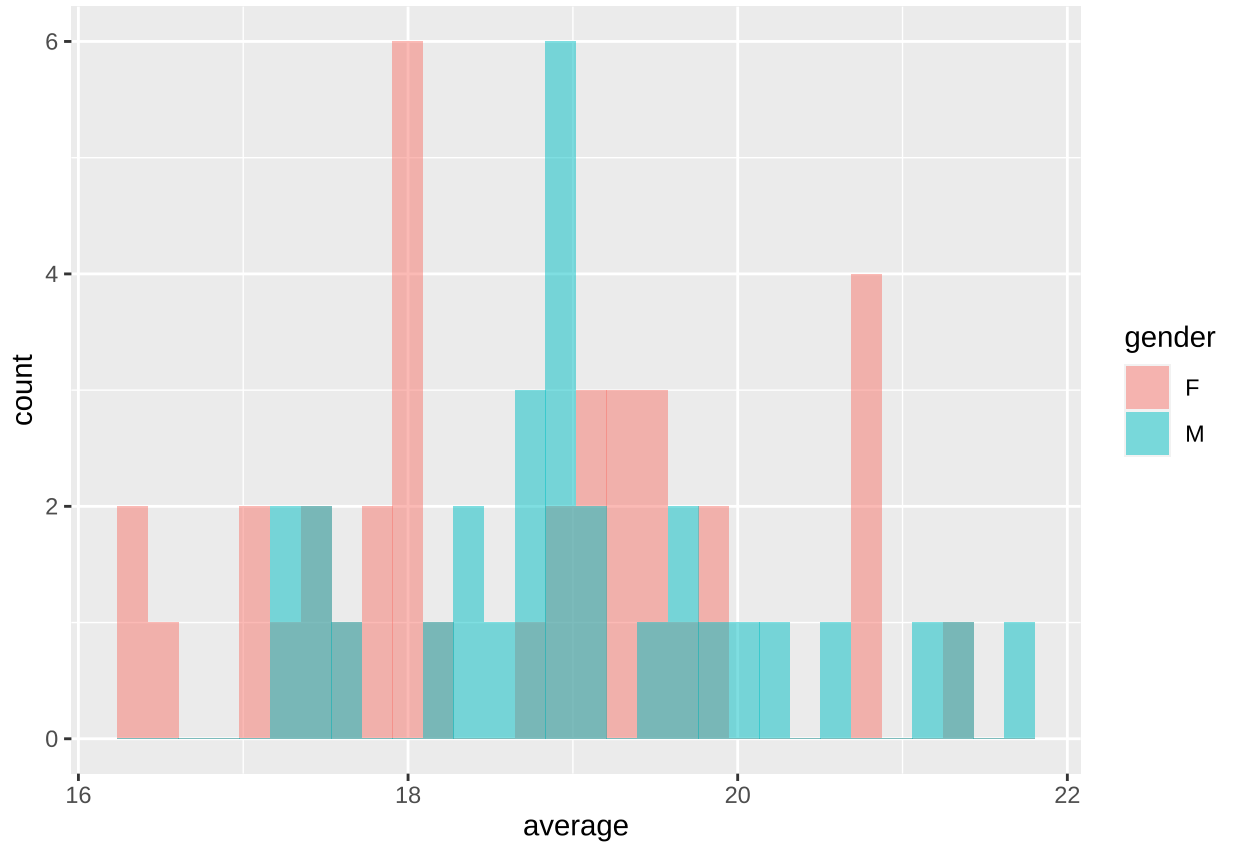


注: NCT的小分队NCT DREAM出道平均年龄为15.6岁。图上只显示NCT全体的平均年龄。

```
ggplot(data = groupdat, aes(x = debut_year, y = average, color = gender)) + geom_point() +
  geom_label(
    label = groupdat$group,
    nudge_x = 0.2,
    nudge_y = 0.1,
    size = 2.5
  ) + style
```




```
ggplot(data = groupdat, aes(x = average, fill = gender)) + geom_histogram(alpha = 0.5, position = "iden
```




```
t.test(dat_f$average, dat_m$average, alternative = "less")
```

```
##
##  Welch Two Sample t-test
##
## data:  dat_f$average and dat_m$average
## t = -1.4122, df = 64.976, p-value = 0.08133
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##        -Inf 0.07747915
## sample estimates:
## mean of x mean of y
##  18.65817  19.08480
```

```
# one-tail test: p-value < 0.1 -> different means

filter(memberdat, gender == "F", age <= 14)
```

```
## # A tibble: 3 x 7
##   group     company  gender show  debut_year member        age
##   <chr>     <chr>    <chr>  <chr>      <int> <chr>       <int>
## 1 New Jeans HYBE     F      N           2022 debut_age5     14
## 2 IZONE     MNET     F      Y           2018 debut_age12    14
## 3 IVE       STARSHIP F      N           2021 debut_age6     14
```

```
filter(memberdat, gender == "M", age <= 14)
```

```
## # A tibble: 1 x 7
##   group company gender show  debut_year member        age
##   <chr> <chr>   <chr>  <chr>      <int> <chr>       <int>
## 1 NCT   SM      M      N           2016 debut_age14    14
```

- Median is the same. The difference between means is statistically significant.

SD of debut ages within groups

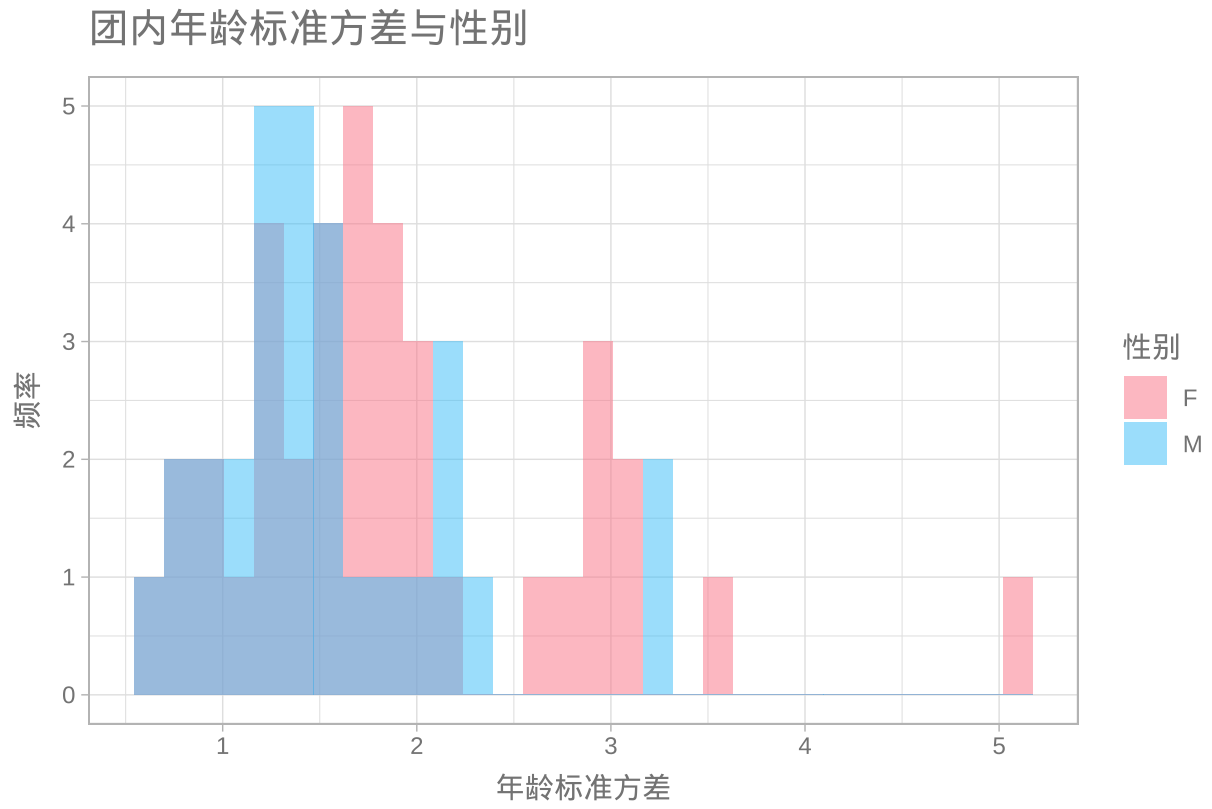
```
median(dat_m$sd)
```

```
## [1] 1.43359
```

```
median(dat_f$sd)
```

```
## [1] 1.722153
```

```
ggplot(data = groupdat, aes(x = sd, fill = gender))+
  geom_histogram(alpha = 0.5, position = "identity")+
  style +
  scale_fill_manual( values = c("#fb7185", "#38bdf8")) +
  labs(title = "      ", x = "    ", y = " ", fill = " ")
```



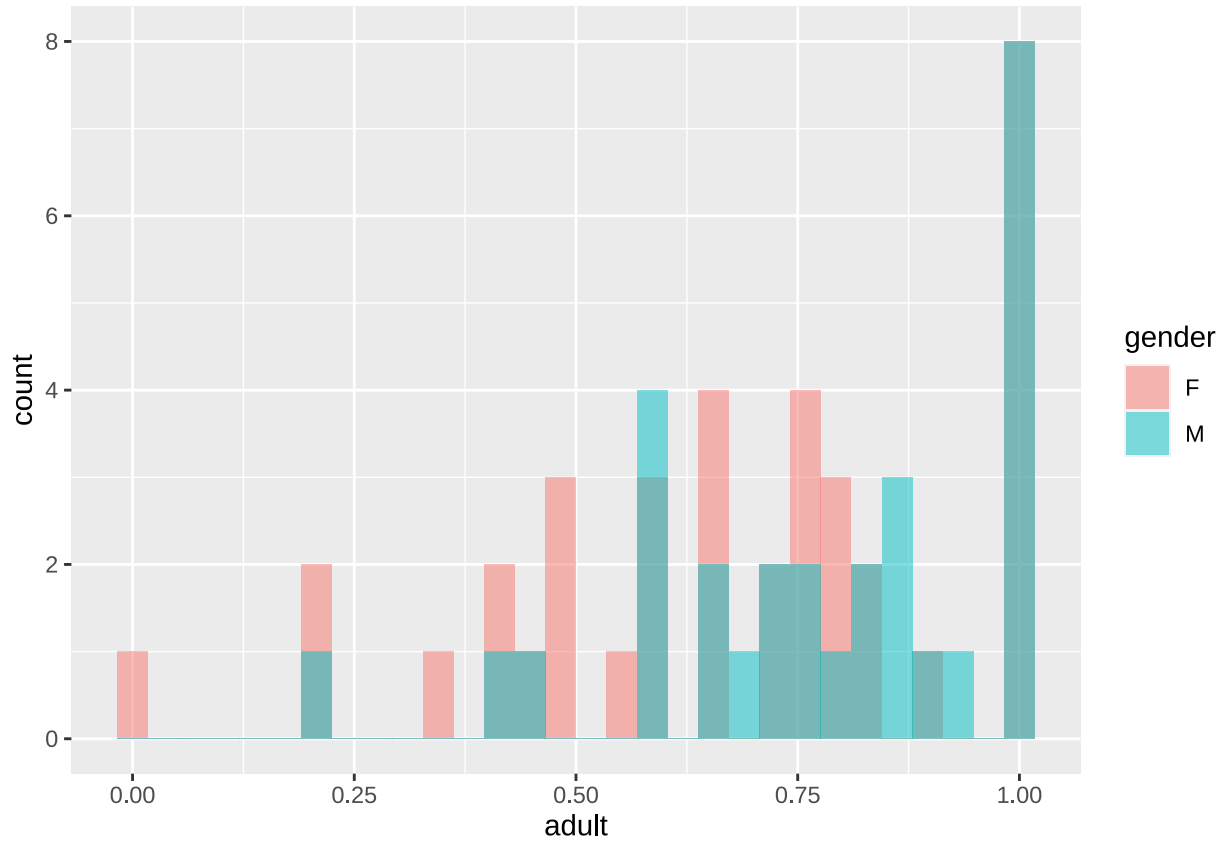
```r
ggsave(
  "stddev_by_gender.png",
  plot = last_plot(),
  device = png)

t.test(dat_m$sd, dat_f$sd, alternative = "less")
```

```
##
##  Welch Two Sample t-test
##
## data:  dat_m$sd and dat_f$sd
## t = -1.7709, df = 65.169, p-value = 0.04063
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##         -Inf -0.01889364
## sample estimates:
## mean of x mean of y
##  1.565520  1.892508
```

Adult rate

```r
ggplot(data = groupdat, aes(x = adult , fill = gender)) +
  geom_histogram(alpha = 0.5, position = "identity")
```



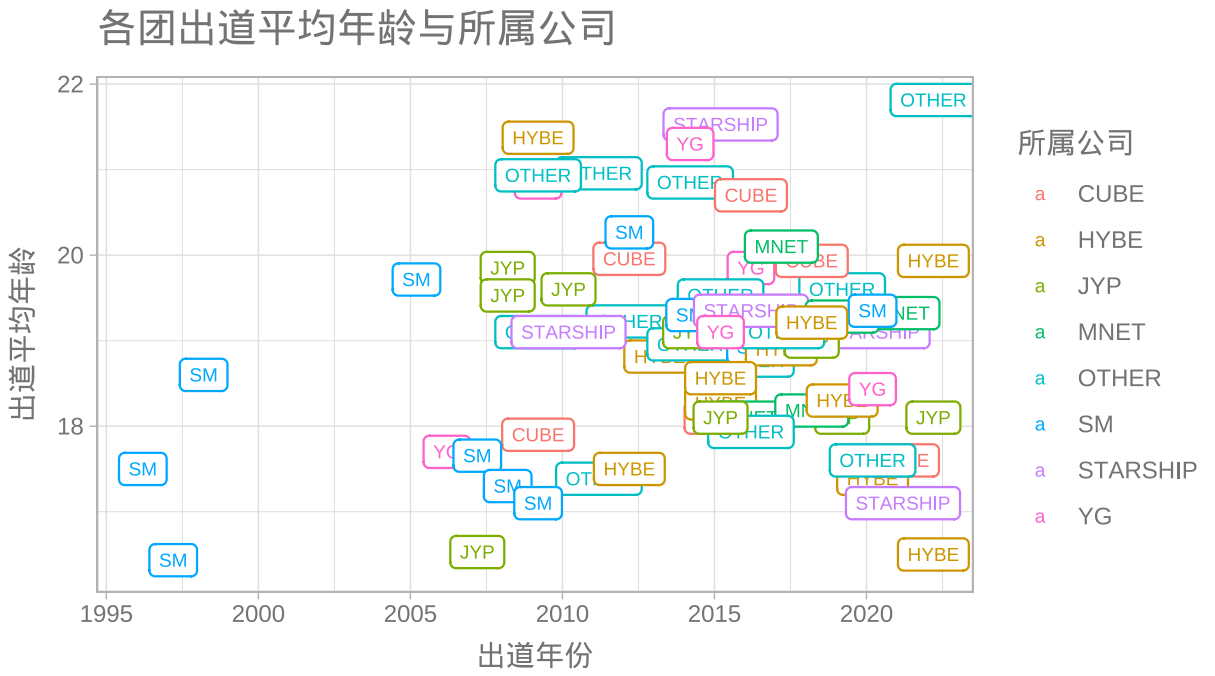
```r
t.test(dat_f$adult, dat_m$adult, alternative = "less")
```

```
##
##  Welch Two Sample t-test
##
## data:  dat_f$adult and dat_m$adult
## t = -1.6752, df = 65.931, p-value = 0.04931
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##           -Inf -0.0003837479
## sample estimates:
## mean of x mean of y
## 0.6817403 0.7742812
```

Analysis by company

Average debut age

```r
ggplot(data = groupdat, aes(x = debut_year, y = average, color = company)) + geom_point() +
  geom_label(
    label = groupdat$company,
    nudge_x = 0.2,
    nudge_y = 0.1,
    label.size = 0.4,
    size = 2.5
  ) + style +
  labs(
    title = "       ",
    x = "  ",
    y = "    ",
    color = "  ",
    caption = " :  HYBE       HYBE "
  )
```



注: 被HYBE收购的公司历代的团体都算在HYBE。

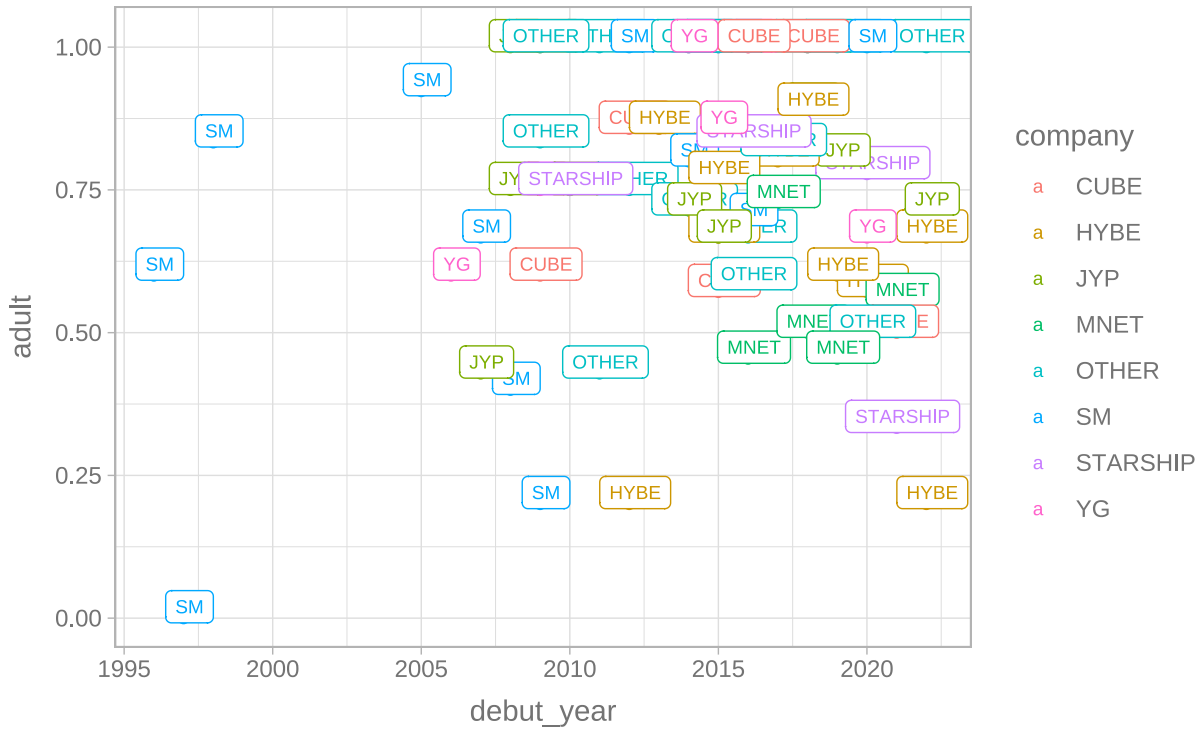
```
ggsave("group_mean_debut_age_by_company.png",
       plot = last_plot(),
       device = png)

summarize(group_by(memberdat, company), avg = mean(age))
```

```
## # A tibble: 8 x 2
##   company    avg
##   <chr>    <dbl>
## 1 CUBE      19.0
## 2 HYBE      18.7
## 3 JYP       18.5
## 4 MNET      18.9
## 5 OTHER     19.2
## 6 SM        18.7
## 7 STARSHIP  19.2
## 8 YG        19.2
```

Adult rate

```
ggplot(data = groupdat, aes(x = debut_year, y = adult, color = company)) + geom_point() +
  geom_label(
    label = groupdat$company,
    nudge_x = 0.2,
    nudge_y = 0.02,
    size = 2.5
  ) + style
```




```
summarize(group_by(groupdat, company),
          total = sum(numMember * adult) / sum(numMember))
```

```
## # A tibble: 8 x 2
##   company  total
##   <chr>    <dbl>
## 1 CUBE     0.767
## 2 HYBE     0.714
## 3 JYP      0.746
## 4 MNET     0.537
## 5 OTHER    0.802
## 6 SM       0.722
## 7 STARSHIP 0.763
## 8 YG       0.784
```

Talent show or not

Average debut age

```
dat_y <- filter(groupdat, show == "Y")
dat_n <- filter(groupdat, show == "N")
median(dat_y$average)
```

```
## [1] 18.59722
```

```r
median(dat_n$average)
```

```
## [1] 19
```

```r
mean(dat_y$average)
```

```
## [1] 18.61052
```

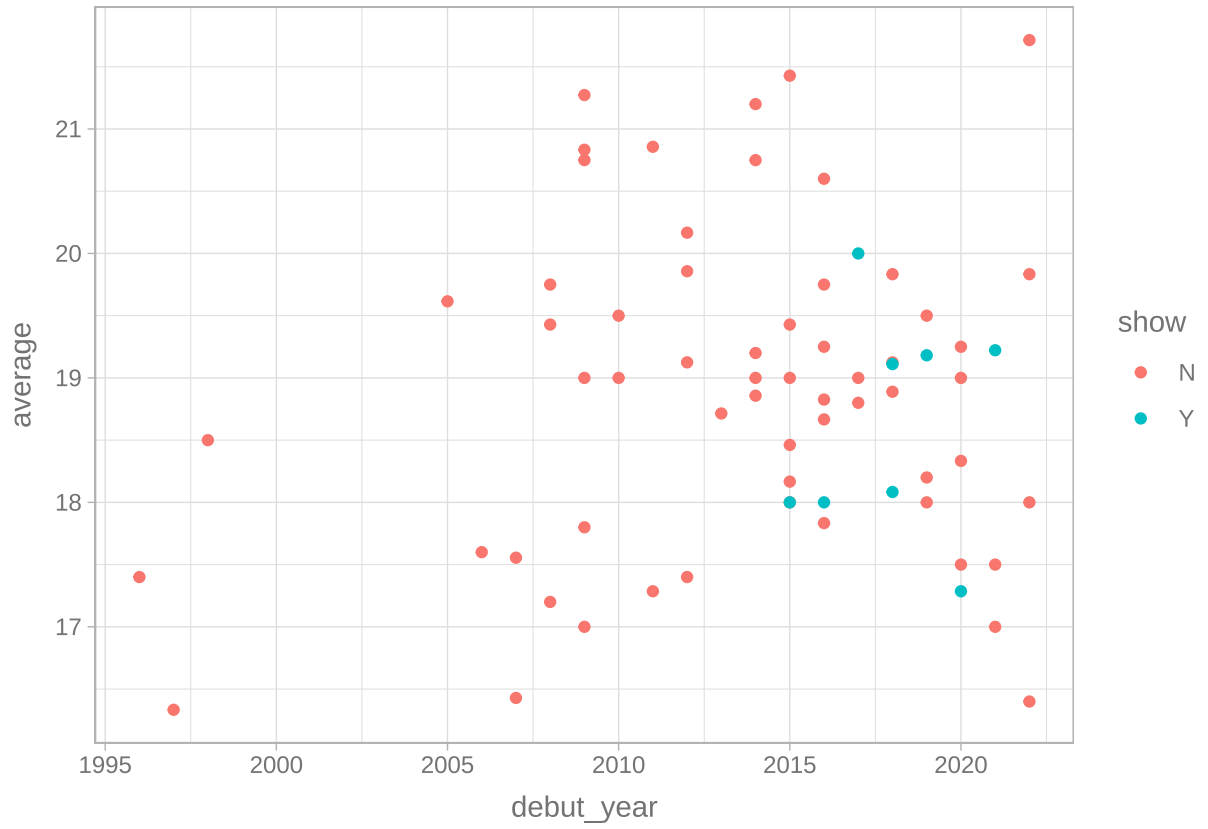
```r
mean(dat_n$average)
```

```
## [1] 18.87784
```

```r
t.test(dat_y$average, dat_m$average, alternative = "less") # p-value is 0.12.
```

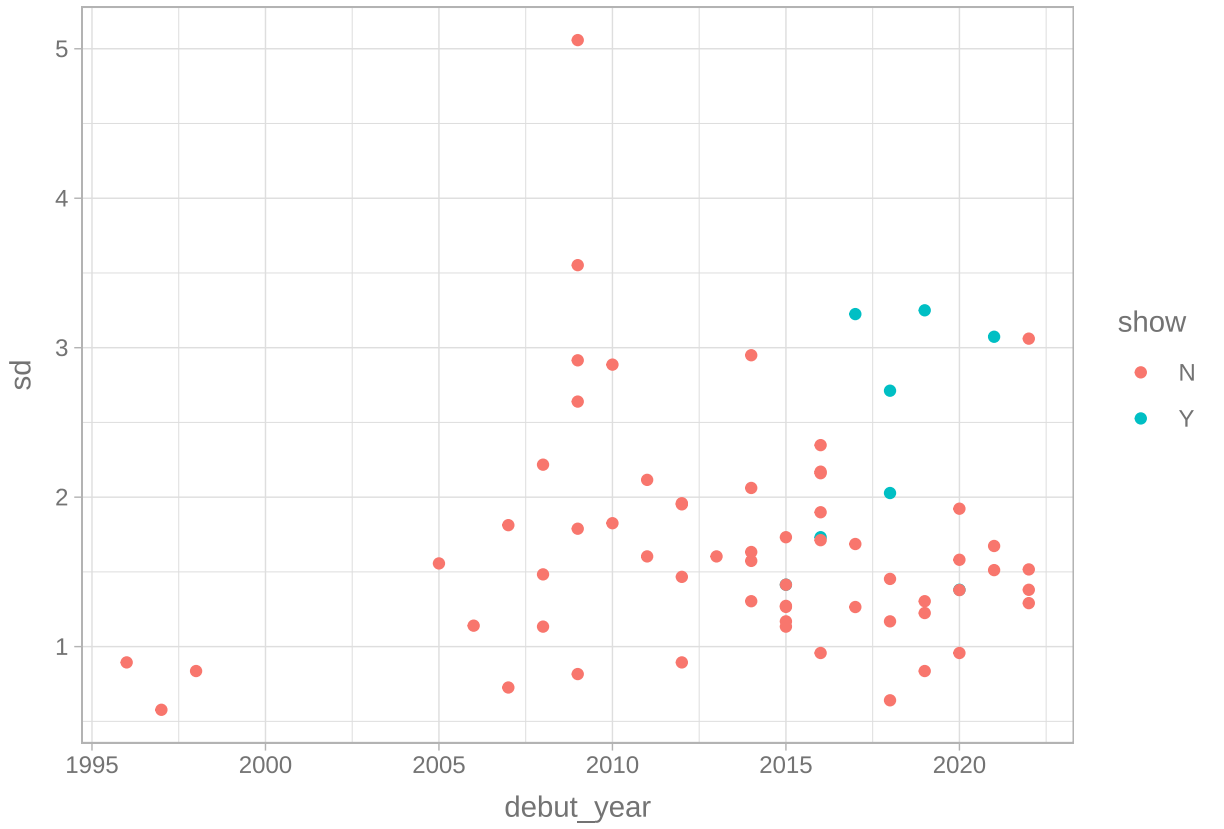
```
##
##  Welch Two Sample t-test
##
## data:  dat_y$average and dat_m$average
## t = -1.2376, df = 14.081, p-value = 0.1181
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf 0.2004423
## sample estimates:
## mean of x mean of y
##  18.61052  19.08480
```

```r
ggplot(data = groupdat, aes(x = debut_year, y = average, color = show)) + geom_point() + style
```



SD

```
ggplot(data = groupdat, aes(x = debut_year, y = sd, color = show)) + geom_point() + style
```



Adult rate

```
ggplot(data = groupdat, aes(x = debut_year, y = adult, color = show)) + geom_point() + style
```

Analysis by generation

- 1st gen: - 2005
- 2nd gen: 2005 - 2009
- 3rd gen: 2010 - 2014
- 4th gen: 2015 - 2018
- 5th gen: 2019 - present

```
groupdat$gen <- NA
groupdat[groupdat$debut_year < 2005, ]$gen <- 1
groupdat[groupdat$debut_year < 2010 & groupdat$debut_year > 2004, ]$gen <- 2
groupdat[groupdat$debut_year < 2015 & groupdat$debut_year > 2009, ]$gen <- 3
groupdat[groupdat$debut_year < 2019 & groupdat$debut_year > 2014, ]$gen <- 4
groupdat[groupdat$debut_year > 2018, ]$gen <- 5

memberdat$gen <- NA
memberdat[memberdat$debut_year < 2005, ]$gen <- 1
memberdat[memberdat$debut_year < 2010 & memberdat$debut_year > 2004, ]$gen <- 2
memberdat[memberdat$debut_year < 2015 & memberdat$debut_year > 2009, ]$gen <- 3
memberdat[memberdat$debut_year < 2019 & memberdat$debut_year > 2014, ]$gen <- 4
memberdat[memberdat$debut_year > 2018, ]$gen <- 5
```

Average debut age

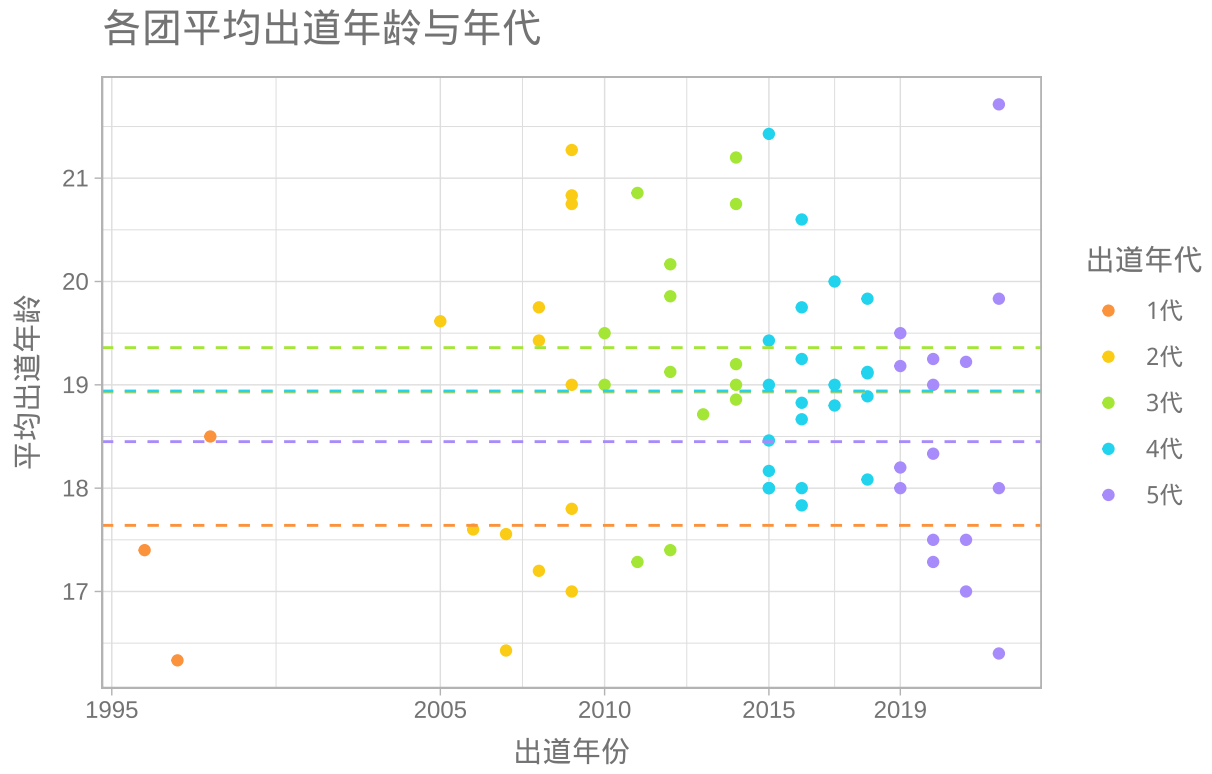
```
summarize(group_by(memberdat, gen), median = median(age))
```

```
## # A tibble: 5 x 2
##     gen median
##   <dbl>  <dbl>
## 1     1     18
## 2     2     19
## 3     3     19
## 4     4     19
## 5     5     18
```

```
summarize(group_by(memberdat, gen), average = mean(age))
```

```
## # A tibble: 5 x 2
##     gen average
##   <dbl>   <dbl>
## 1     1    17.6
## 2     2    18.9
## 3     3    19.4
## 4     4    18.9
## 5     5    18.6
```

```
ggplot(data = groupdat, aes(x = debut_year, y = average, color = factor(gen))) +
  geom_point(alpha = 1, size = 1.5) + style +
  scale_x_continuous(breaks = c(1995, 2005, 2010, 2015, 2019)) +
  labs(
    title = "      ",
    y = "   ",
    x = "  ",
    color = "  ",
    caption = " : 2 4      "
  )  +
  scale_color_manual(
    values = c("#fb923c", "#facc15", "#a3e635", "#22d3ee", "#a78bfa"),
    labels = c("1 ", "2 ", "3 ", "4 ", "5 ")
  )  +
  geom_hline(
    yintercept = c(17.64, 18.93, 19.36, 18.94, 18.45),
    linetype = "dashed",
    size = 0.5,
    color = c("#fb923c", "#facc15", "#a3e635", "#22d3ee", "#a78bfa")
  )
```



注: 2代与4代平均出道年龄线重合。

```
ggsave(
  "group_mean_debut_age_by_generation.png",
  plot = last_plot(),
  device = png)

four <- filter(memberdat, gen == 4)
five <- filter(memberdat, gen == 5)
t.test(four$age, five$age, alternative = "greater")
```

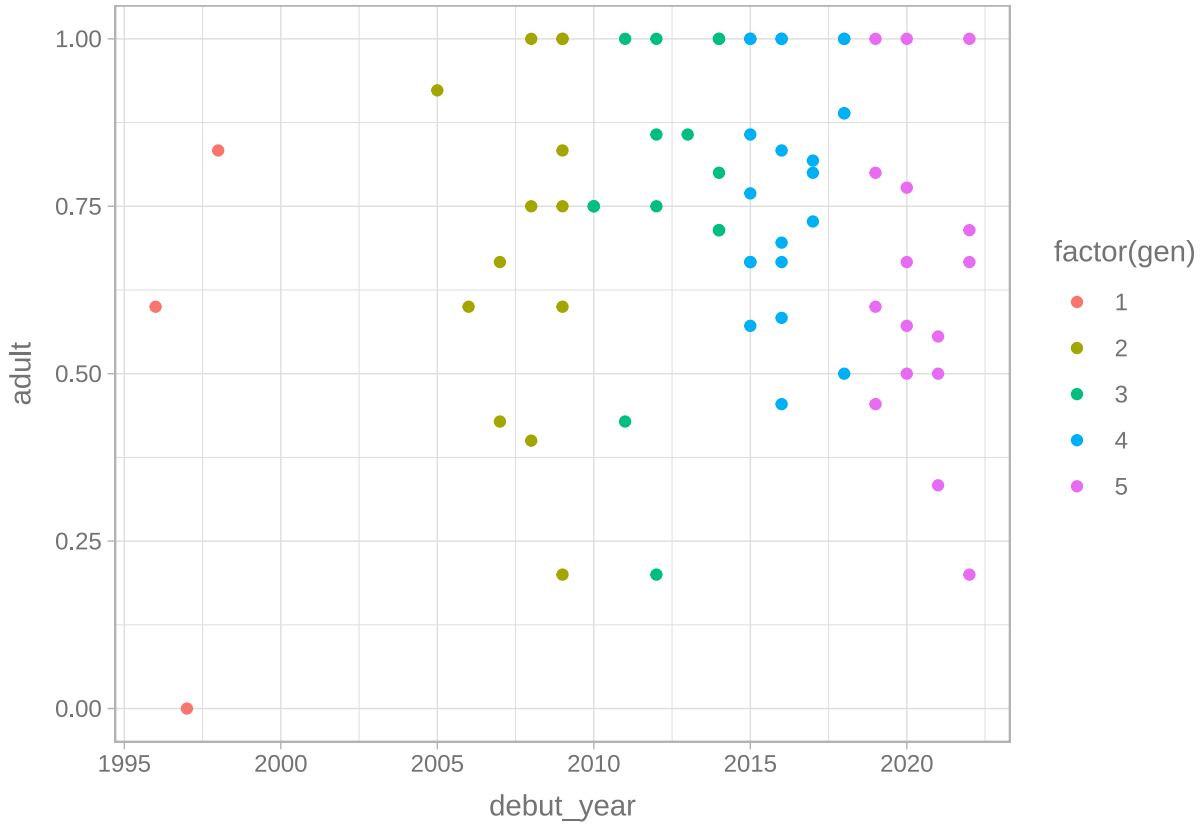
```
##
##  Welch Two Sample t-test
##
## data:  four$age and five$age
## t = 1.5221, df = 207.03, p-value = 0.06476
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -0.03254919         Inf
## sample estimates:
## mean of x mean of y
##  18.93810  18.55752
```

Adult rate

```
summarize(group_by(groupdat, gen),
          adult = sum(numMember * adult) / sum(numMember))
```

```
## # A tibble: 5 x 2
##     gen adult
##   <dbl> <dbl>
## 1     1 0.571
## 2     2 0.741
## 3     3 0.787
## 4     4 0.767
## 5     5 0.637
```

```
ggplot(data = groupdat, aes(x = debut_year, y = adult, color = factor(gen))) + geom_point() + style
```



Analysis by company and gender

```
sum_avg <- summarize(group_by(groupdat, company, gender), average = sum(ageSum)/sum(numMember))
sum_avg
```

```
## # A tibble: 16 x 3
## # Groups:   company [8]
```

```
##    company  gender average
##    <chr>    <chr>    <dbl>
##  1 CUBE     F         18.2
##  2 CUBE     M         20.3
##  3 HYBE     F         19.2
##  4 HYBE     M         18.1
##  5 JYP      F         17.8
##  6 JYP      M         19.2
##  7 MNET     F         18.4
##  8 MNET     M         19.6
##  9 OTHER    F         19.0
## 10 OTHER    M         19.5
## 11 SM       F         17.9
## 12 SM       M         19.0
## 13 STARSHIP F         18.6
## 14 STARSHIP M         20.1
## 15 YG       F         20.2
## 16 YG       M         18.9
```

```
summarize(group_by(groupdat, company, gender),  median = median(average))
```

```
## # A tibble: 16 x 3
## # Groups:   company [8]
##    company  gender median
##    <chr>    <chr>   <dbl>
##  1 CUBE     F        17.9
##  2 CUBE     M        20.2
##  3 HYBE     F        19.0
##  4 HYBE     M        18.2
##  5 JYP      F        18
##  6 JYP      M        19.2
##  7 MNET     F        18.1
##  8 MNET     M        19.6
##  9 OTHER    F        19.3
## 10 OTHER    M        19
## 11 SM       F        17.6
## 12 SM       M        18.7
## 13 STARSHIP F        19
## 14 STARSHIP M        20.2
## 15 YG       F        20.2
## 16 YG       M        18.7
```

```
sum_adult <- summarize(group_by(groupdat, company, gender),adult=sum(numMember*adult)/sum(numMember))
```

1. average by company by gender

- CUBE: male(20.3) - female(18.2) = 2.1
- HYBE: male(18.1) - female(19.2) = -1.1
- JYP: male(19.2) - female(17.8) = 1.4
- OTHER: 0.5

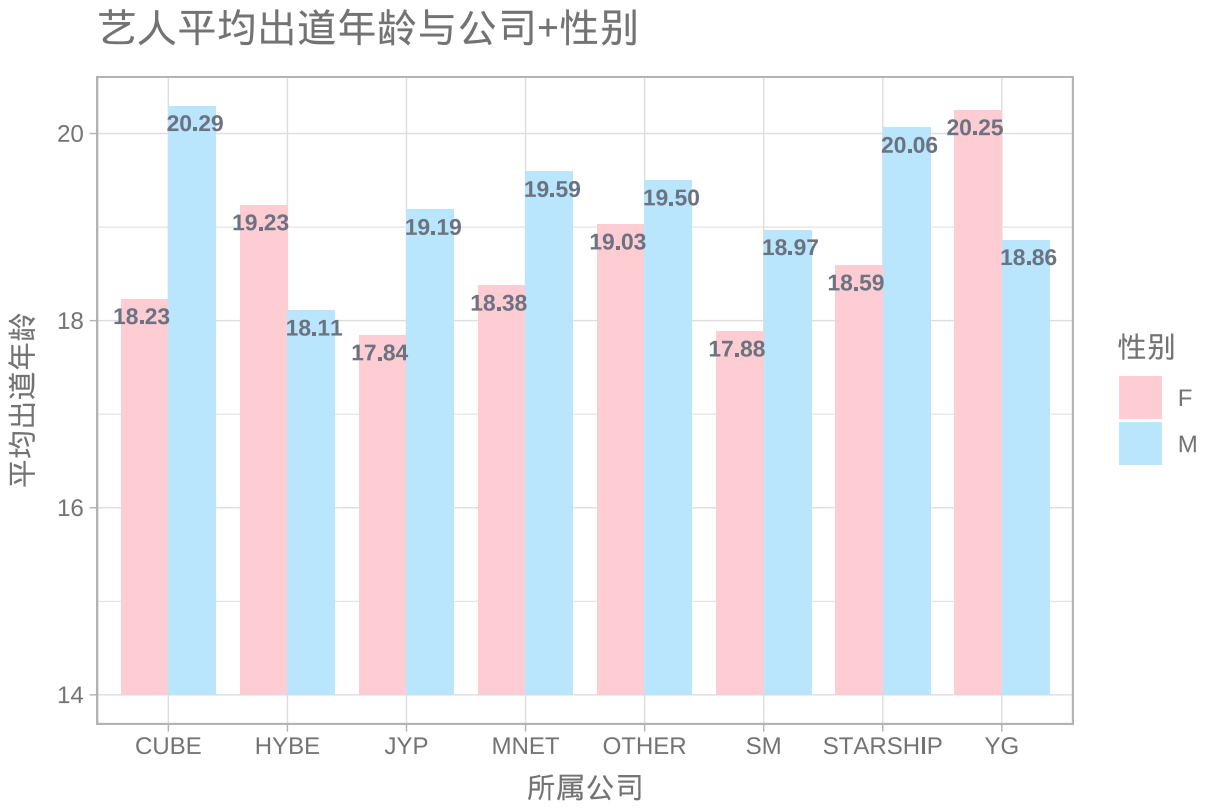
2. median

- CUBE: male(22.2) - female(17.9) = 2.3

3. adult%

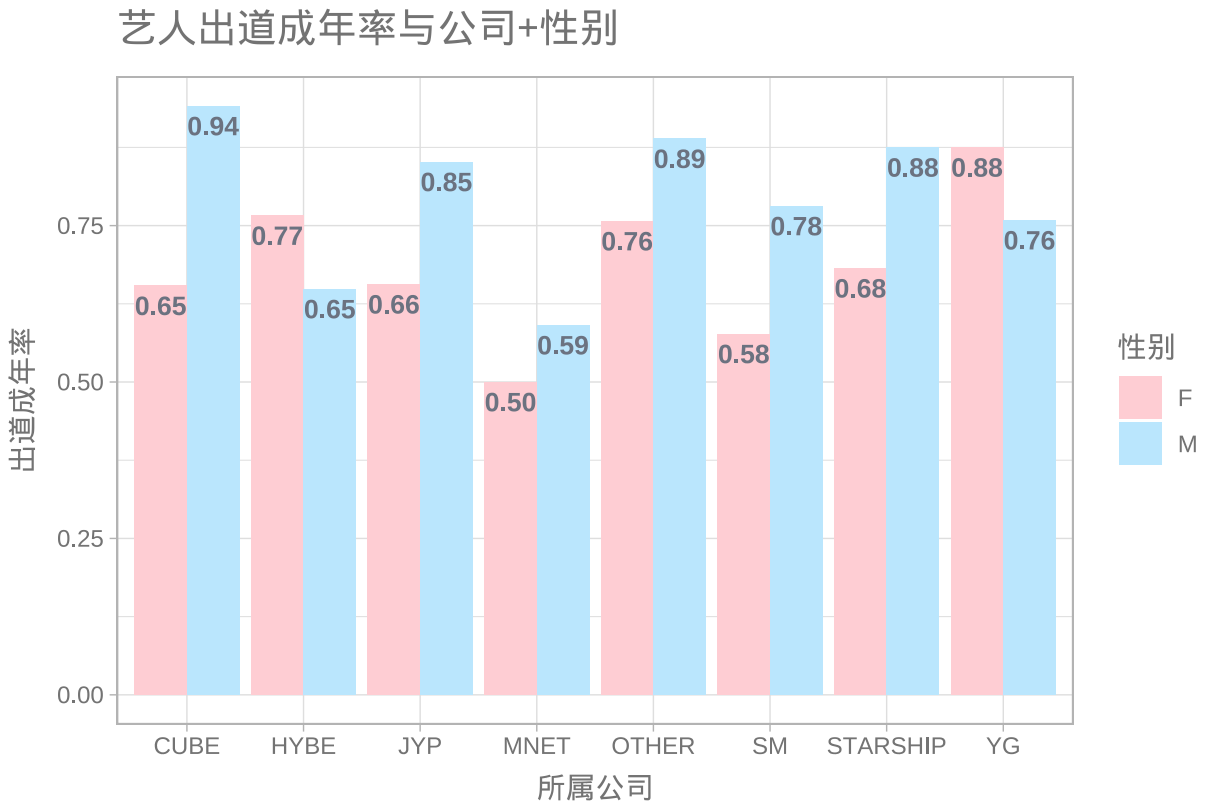
- CUBE: male(0.94) - female(0.65) = 0.29!!!
- HYBE: female > male
- JYP: male(0.85) - female(0.66) = 0.19
- OTHER: female < male

```
ggplot(data = sum_avg, aes(x = company, y = average - 14, fill = gender)) +
  geom_bar(stat = "identity",
           position = position_dodge(),
           width = 0.8) +
  scale_fill_manual(values = c('#fecdd3', '#bae6fd')) + style +
  geom_text(
    aes(label = format(round(average, 2), nsmall = 2)),
    vjust = 1.6,
    color = "#6b7280",
    position = position_dodge(0.9),
    size = 3,
    fontface = "bold"
  ) +
  labs(title = "      + ",
       x = "  ",
       y = "    ",
       fill = " ") +
  scale_y_continuous(labels = c("14", "16", "18", "20"),
                     breaks = c(0, 2, 4, 6))
```




```
ggsave(
  "mean_debut_age_by_company_and_gender.png",
  plot = last_plot(),
  device = png)
```

```
ggplot(data=sum_adult, aes(x = company, y = adult, fill=gender)) +
geom_bar(stat="identity", position=position_dodge()) +
scale_fill_manual(values=c('#fecdd3','#bae6fd')) + style +
    geom_text(aes(label= format(round(adult,2), nsmall = 2)), vjust=1.6, color="#6b7280",
            position = position_dodge(0.9), size=3.5, fontface = "bold") +
  labs(title = "      + ", x = "  ", y = "   ", fill = " ")
```




```
ggsave(
  "adult_percentage_by_company_and_gender.png",
  plot = last_plot(),
  device = png)
```

Analysis by gender and generation

```
sum_avg <- summarize(group_by(memberdat, gen, gender), average = mean(age), count = n())
sum_adult <- summarize(group_by(groupdat, gen, gender),  adult = sum(numMember*adult) / sum(numMember))
sum_avg
```

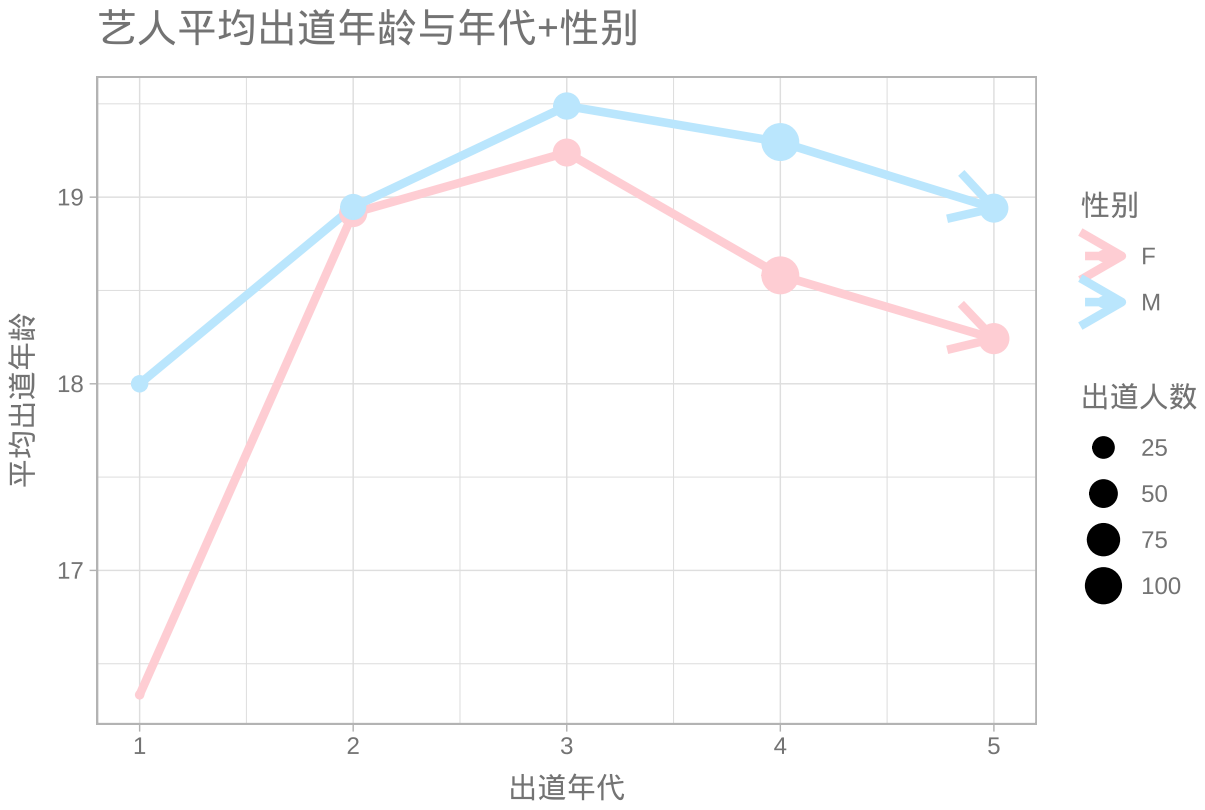
```
## # A tibble: 10 x 4
## # Groups:   gen [5]
##      gen gender average count
##    <dbl> <chr>    <dbl> <int>
##  1     1 F         16.3     3
##  2     1 M         18      11
##  3     2 F         18.9    47
##  4     2 M         18.9    38
##  5     3 F         19.2    46
##  6     3 M         19.5    43
##  7     4 F         18.6   105
```

```
##  8     4 M         19.3   105
##  9     5 F         18.2    62
## 10     5 M         18.9    51
```

```
sum_adult
```

```
## # A tibble: 10 x 3
## # Groups:   gen [5]
##      gen gender adult
##    <dbl> <chr>  <dbl>
##  1     1 F      0
##  2     1 M      0.727
##  3     2 F      0.681
##  4     2 M      0.816
##  5     3 F      0.761
##  6     3 M      0.814
##  7     4 F      0.714
##  8     4 M      0.819
##  9     5 F      0.613
## 10     5 M      0.667
```

```
ggplot(data = sum_avg) + geom_path(aes(x = gen, y = average, color = gender),
                                   size = 1.5,
                                   arrow = arrow()) + geom_point(aes(
                                     x = gen,
                                     y = average,
                                     color = gender,
                                     size = count
                                   )) + style +
  scale_color_manual(values = c('#fecdd3', '#bae6fd')) +
  labs(
    title = "      + ",
    x = "  ",
    y = "    ",
    color = " ",
    size = "  "
  )
```




```
ggsave(
  "mean_debut_age_by_generation_and_gender.png",
  plot = last_plot(),
  device = png)

mean(filter(memberdat,gender=="F",gen==4)$age) - mean(filter(memberdat,gender=="M",gen==4)$age)
```

```
## [1] -0.7142857
```