DA_HW1

Na SeungChan

2023-10-22

Problem 1

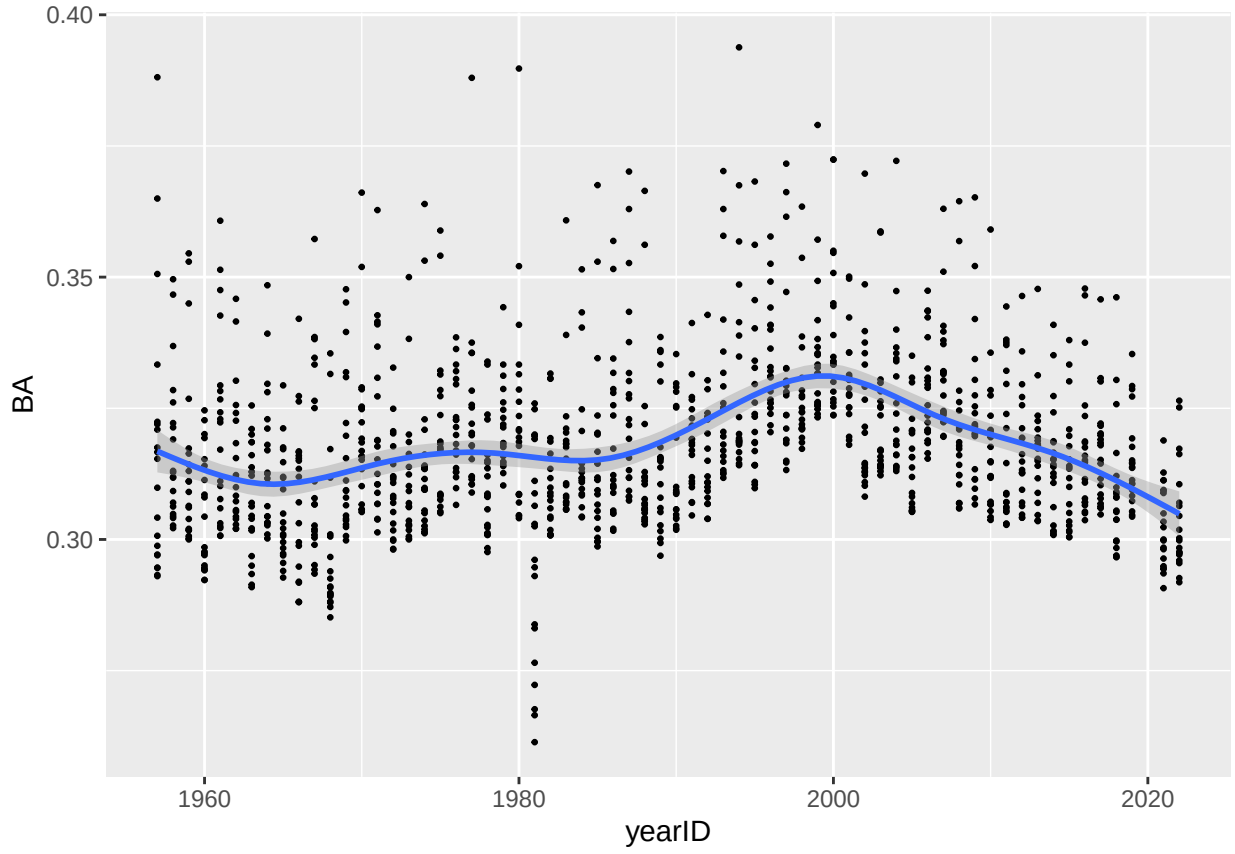
```
df_BA <- Batting %>%
  filter(AB >= 400, yearID >= 1957) %>%
  mutate(BA = H/AB)
```

(a)

```
df_q1a <- df_BA %>%
  group_by(yearID) %>%
  arrange(desc(BA)) %>%
  slice_head(n = 20)

ggplot(data = df_q1a) +
  geom_point(mapping = aes(x = yearID, y = BA), size = 0.5) +
  geom_smooth(mapping = aes(x = yearID, y = BA))
```

```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```




```
#ggplot(data = df_q1a, mapping = aes(x = yearID, y = BA)) +
#  geom_point() +
#  geom_smooth(mapping = aes(yearID ~ BA))
```

(b)

```
df_q1b <- df_BA %>%
  group_by(yearID) %>%
  arrange(desc(BA)) %>%
  slice_head(n = 5) %>%
  ungroup() %>%
  left_join(People) %>%
  mutate(BA = round(BA, 3), fullName = paste(nameGiven, nameLast)) %>%
  select(fullName, yearID, BA)
```

```
## Joining with `by = join_by(playerID)`
```

```
df_q1b
```

```
## # A tibble: 325 x 3
##    fullName                yearID    BA
##    <chr>                    <int> <dbl>
```

```
##  1 Theodore Samuel Williams   1957 0.388
##  2 Mickey Charles Mantle      1957 0.365
##  3 Stanley Frank Musial       1957 0.351
##  4 Willie Howard Mays         1957 0.333
##  5 Frank Robinson             1957 0.322
##  6 Don Richard Ashburn        1958 0.35
##  7 Willie Howard Mays         1958 0.347
##  8 Stanley Frank Musial       1958 0.337
##  9 Theodore Samuel Williams   1958 0.328
## 10 Henry Louis Aaron          1958 0.326
## # i 315 more rows
```

(c)

```
df_q1c <- df_BA %>%
  filter(playerID == 'gwynnto01') %>%
  slice_max(order_by = BA, n = 3) %>%
  select(yearID, BA)

pluck(df_q1c[2], 1)
```

```
## [1] 0.3937947 0.3716216 0.3701188
```

(d)

```
df_q1d <- df_BA %>%
  group_by(yearID, lgID) %>%
  slice_max(order_by = BA, n = 1) %>%
  ungroup() %>%
  filter(playerID == 'gwynnto01')

pluck(df_q1d[2], 1)
```

```
## [1] 1984 1987 1988 1989 1994 1995 1996 1997
```

(e)

```
df_q1e <- df_BA %>%
  filter(yearID >= 2001) %>%
  group_by(lgID) %>%
  arrange(desc(BA)) %>%
  slice_head(n = 1) %>%
  ungroup() %>%
  left_join(People) %>%
  mutate(fullName = paste(nameGiven, nameLast)) %>%
  select(fullName, lgID, yearID, BA)
```

```
## Joining with `by = join_by(playerID)`
```

```
df_q1e
```

```
## # A tibble: 2 x 4
##   fullName          lgID  yearID    BA
##   <chr>             <fct>  <int> <dbl>
## 1 Ichiro Suzuki     AL      2004 0.372
## 2 Barry Lamar Bonds NL      2002 0.370
```

(f)

```
df_q1f <- df_BA %>%
  filter(yearID == 2021, lgID == 'NL', BA >= 0.3) %>%
  arrange(desc(BA)) %>%
  left_join(People) %>%
  mutate(fullName = paste(nameGiven, nameLast)) %>%
  select(fullName, lgID, yearID, BA)
```

```
## Joining with `by = join_by(playerID)`
```

```
pluck(df_q1f[1], 1)[1]
```

```
## [1] "Juan Jose Soto"
```

(g)

```
df_q1g <- Batting %>%
  filter(playerID == 'turnetr01', yearID == 2021) %>%
  group_by(playerID) %>%
  summarise(tH = sum(H), tBA = sum(AB)) %>%
  ungroup() %>%
  mutate(tAB = tH/tBA)

df_q1g
```

```
## # A tibble: 1 x 4
##   playerID     tH   tBA   tAB
##   <chr>     <int> <int> <dbl>
## 1 turnetr01   195   595 0.328
```

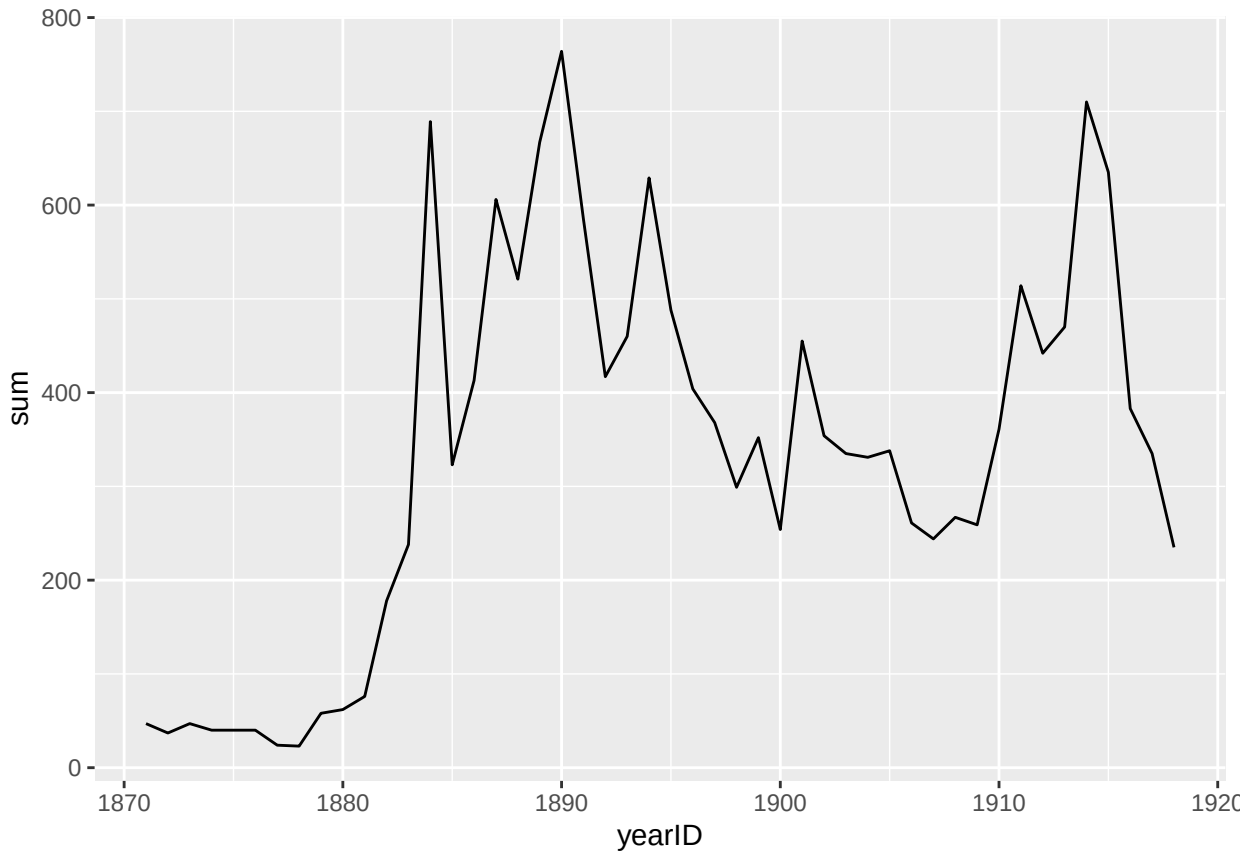
Problem 2

(a)

```
df_q2a <- Batting %>%
  group_by(yearID) %>%
  summarise(sum = sum(HR))
df_q2a
```

```
## # A tibble: 152 x 2
##    yearID   sum
##     <int> <int>
##  1   1871    47
##  2   1872    37
##  3   1873    47
##  4   1874    40
##  5   1875    40
##  6   1876    40
##  7   1877    24
##  8   1878    23
##  9   1879    58
## 10   1880    62
## # i 142 more rows
```

```
ggplot(data = df_q2a %>% filter(yearID <= 1918)) +
  geom_line(mapping = aes(x = yearID, y = sum))
```



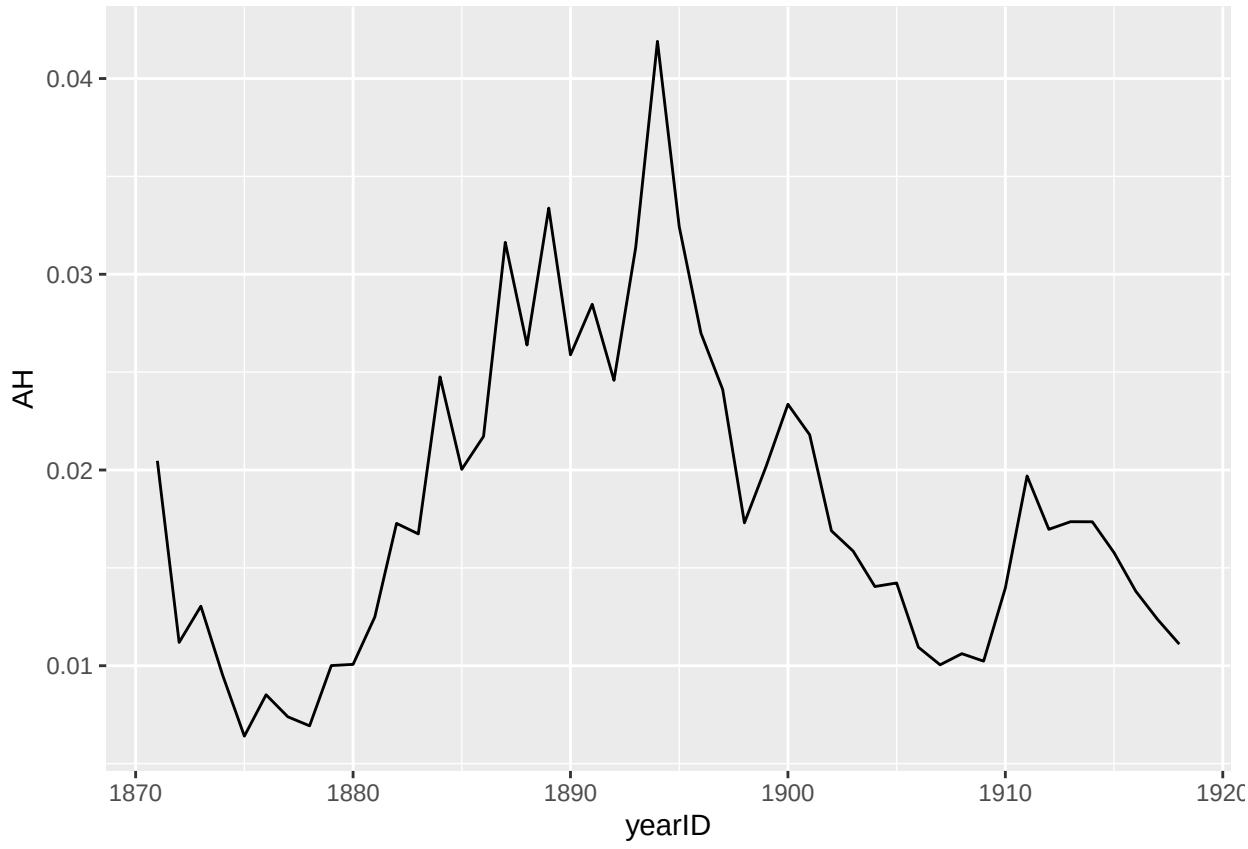
1880년대 초반부터 total HR 크게 증가. 그러나 이후의 명확한 증가 추세가 유지되지는 않았음.

(b)

```
df_q2b <- Batting %>%
  group_by(yearID) %>%
  summarise(tH = sum(HR), tG = sum(G), AH = tH/tG)
df_q2b
```

```
## # A tibble: 152 x 4
##    yearID    tH    tG      AH
##     <int> <int> <int>   <dbl>
##  1   1871    47  2296 0.0205
##  2   1872    37  3306 0.0112
##  3   1873    47  3604 0.0130
##  4   1874    40  4199 0.00953
##  5   1875    40  6248 0.00640
##  6   1876    40  4696 0.00852
##  7   1877    24  3247 0.00739
##  8   1878    23  3319 0.00693
##  9   1879    58  5795 0.0100
## 10   1880    62  6157 0.0101
## # i 142 more rows
```

```
ggplot(data = df_q2b %>% filter(yearID <= 1918)) +
  geom_line(mapping = aes(x = yearID, y = AH))
```



앞서 문제와 같이 1880년대 초반부터 홈런 수가 증가한 경향이 존재한 것은 사실이나, '이후의 명확한 증가 추세가 유지되지는 않음' 부분은 과거와 유사한 수준의 홈런으로 돌아옴. 즉, (a)에서 해당 시기 홈런 개수가 증가한 것은 단순 경기수의 증가 영향에 가까워 보임.

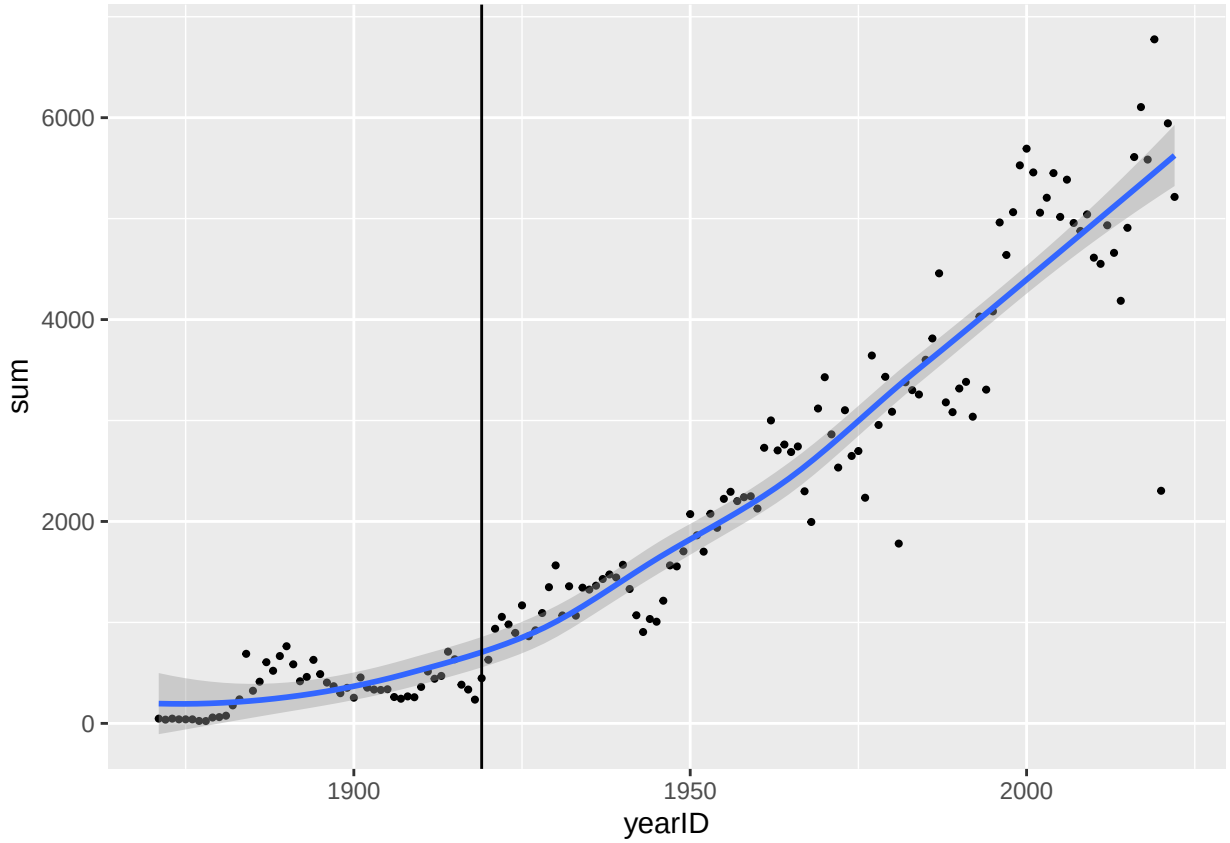
(c)

```
df_q2c <- Batting %>%
  group_by(yearID) %>%
  summarise(sum = sum(HR))
df_q2c
```

```
## # A tibble: 152 x 2
##    yearID   sum
##     <int> <int>
##  1   1871    47
##  2   1872    37
##  3   1873    47
##  4   1874    40
##  5   1875    40
##  6   1876    40
##  7   1877    24
##  8   1878    23
##  9   1879    58
## 10   1880    62
## # i 142 more rows
```

```
ggplot(data = df_q2c, mapping = aes(x = yearID, y = sum)) +
  geom_point(size = 0.75) +
  geom_smooth() +
  geom_vline(xintercept = 1919)
```

```
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```



지속적으로 증가하고, 2000년대 부근에 두드러지는 봉이 존재하는 정도이다. 2020년 단축 시즌으로 인해 크게 낮아진 점 역시 보인다.

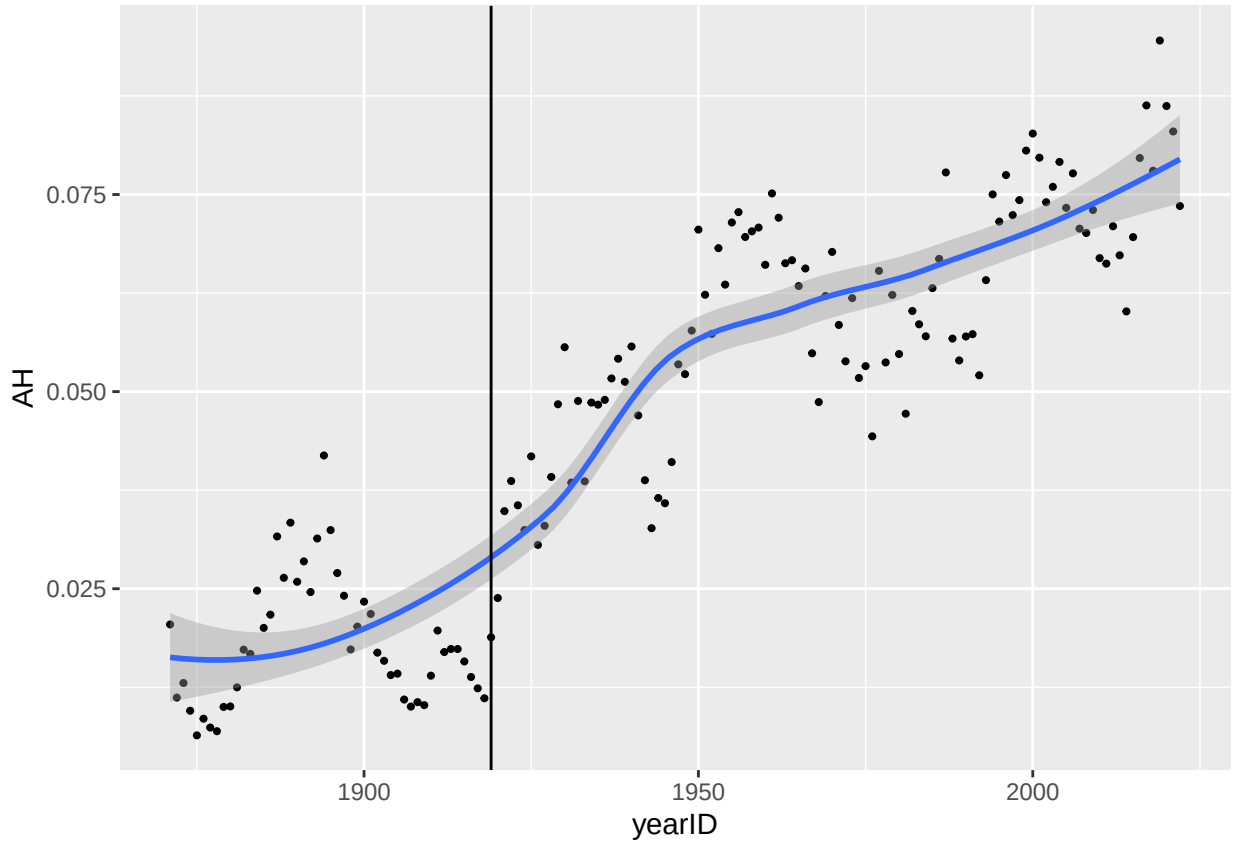
(d)

```
df_q2d <- Batting %>%
  group_by(yearID) %>%
  summarise(tH = sum(HR), tG = sum(G), AH = tH/tG)
df_q2d
```

```
## # A tibble: 152 x 4
##    yearID    tH    tG      AH
##     <int> <int> <int>   <dbl>
##  1   1871    47  2296 0.0205
##  2   1872    37  3306 0.0112
##  3   1873    47  3604 0.0130
##  4   1874    40  4199 0.00953
##  5   1875    40  6248 0.00640
##  6   1876    40  4696 0.00852
##  7   1877    24  3247 0.00739
##  8   1878    23  3319 0.00693
##  9   1879    58  5795 0.0100
## 10   1880    62  6157 0.0101
## # i 142 more rows
```

```
ggplot(data = df_q2d) +
  geom_point(mapping = aes(x = yearID, y = AH), size = 0.75) +
  geom_smooth(mapping = aes(x = yearID, y = AH)) +
  geom_vline(xintercept = 1919)
```

```
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```



1919년을 기점으로 크게 증가하는 경향 존재. 특히 1950년대 부근, 2000년대 부근, 최근 등 비율이 크게 증가하는 등 전반적으로 비율로 계산하는 경우 불안정한 경향이 존재함.