EDA

Group E

12/14/2021

Loading Data

```
wtaresults <- data.frame(matrix(ncol = 49, nrow = 0))
for(i in 1968:2021) {
  temp <- read_csv(paste0("https://raw.githubusercontent.com/JeffSackmann/tennis_wta/master/wta_matches_
  wtaresults <- rbind(wtaresults, temp)
}
```

```
atpresults <- data.frame(matrix(ncol = 49, nrow = 0))
for(i in 1968:2021) {
  temp <- read_csv(paste0("https://raw.githubusercontent.com/JeffSackmann/tennis_atp/master/atp_matches_
  atpresults <- rbind(atpresults, temp)
}
```

```
library(tidyverse)
# adding column to prepare to combine datasets
wtaresults <- wtaresults %>%
  mutate(tour = "WTA")

atpresults <- atpresults %>%
  mutate(tour = "ATP")

# moving tour column to front for ease
wtaresults <- wtaresults[,c(50,1:49)]
atpresults <- atpresults[,c(50,1:49)]

# combining the datasets
tennis_results <- rbind(wtaresults, atpresults)

# making date objects from date
library(lubridate)
tennis_results <- tennis_results %>%
  mutate(tourney_date = ymd(tourney_date)) %>%
  mutate(year = year(tourney_date)) %>%
  mutate(month = month(tourney_date))

# reorganizing date columns together
tennis_results <- tennis_results[,c(1:7,51:52,8:50)]
```

Data Cleaning

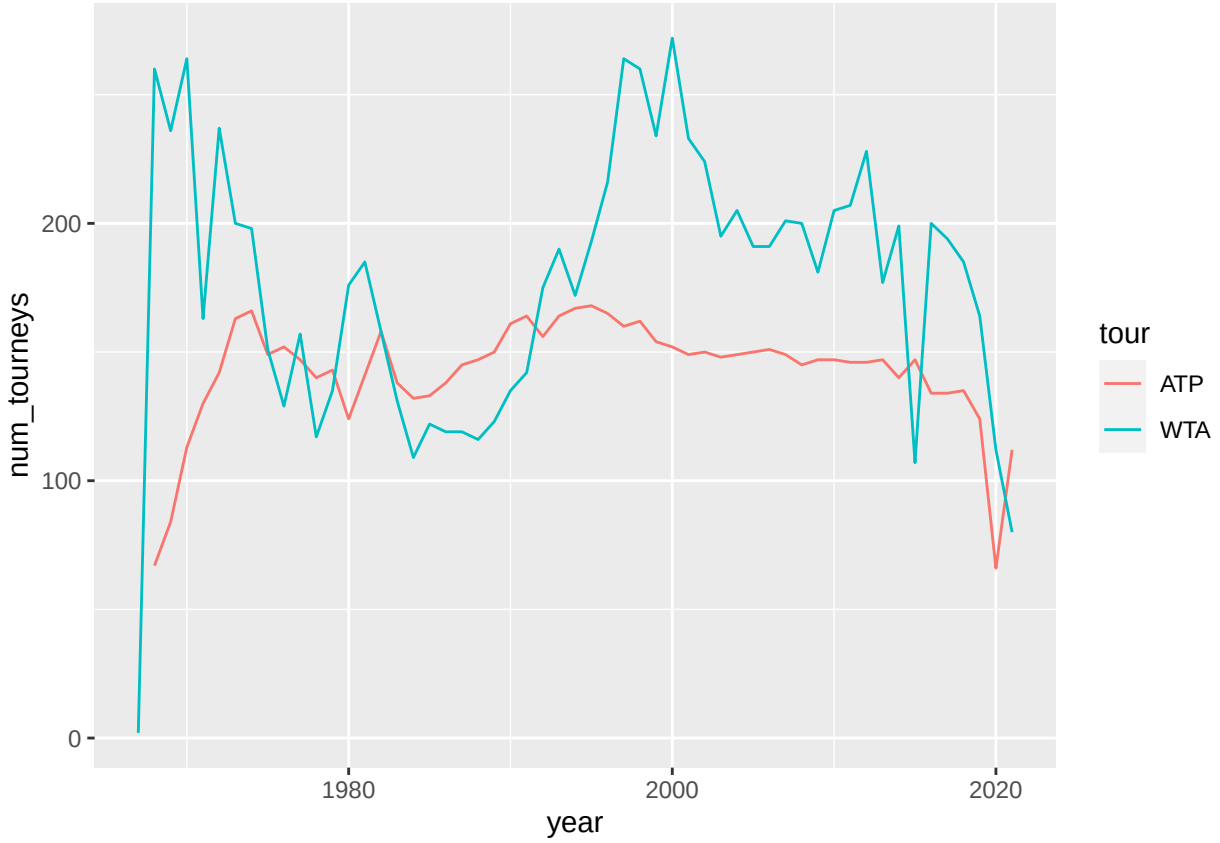
```
# making the alternate abbreviation consistent
tennis_results$winner_entry[tennis_results$winner_entry=="Alt"] <- "ALT"

# ensuring that `winner_seed` is of numeric type instead of character
tennis_results <- tennis_results %>%
  mutate(winner_seed = as.integer(winner_seed)) %>%
  mutate(loser_seed = as.integer(loser_seed))

# fixing `loser_entry` typos
tennis_results <- tennis_results %>%
  mutate(loser_entry = case_when(
    loser_entry == 'A' ~ 'ALT',
    loser_entry == 'Alt' ~ 'ALT',
    loser_entry == 'wc' ~ 'WC',
    loser_entry == 'S' ~ 'SE',
    TRUE ~ loser_entry
  ))

# cleaning datasets with duplicated information
clean_1973_surbiton <- tennis_results %>%
  filter(tourney_id == '1973-1098',
         match_num != 32,
         match_num != 33)

clean_1981_johannesburg <- tennis_results %>%
  filter(tourney_id == '1981-1099') %>%
  slice_tail(n = 11)

clean_1990_taranto <- tennis_results %>%
  filter(tourney_id == '1990-W-WT-ITA-01A-1990',
         !(match_num == 29 & round == 'R32'),
         !(match_num == 30 & round == 'R32'),
         !(match_num == 31 & round == 'R32'),
         match_num <= 31)

clean_1991_stpetersburg <- tennis_results %>%
  filter(tourney_id == '1991-W-WT-URS-01A-1991',
         !(match_num == 29 & round == 'R32'),
         !(match_num == 30 & round == 'R32'),
         !(match_num == 31 & round == 'R32'),
         match_num <= 31)

clean_1991_oakland <- tennis_results %>%
  filter(tourney_id == '1991-W-WT-USA-19A-1991') %>%
  slice_head(n = 27)

clean_1992_oklahoma <- tennis_results %>%
  filter(tourney_id == '1992-W-WT-USA-02A-1992',
         !(match_num == 28 & round == 'R32'),
         !(match_num == 29 & round == 'R32'),
         !(match_num == 30 & round == 'R32'),
```

```
         !(match_num == 31 & round == 'R32'),
         match_num <= 31)

# vector of RR tourneys or tourneys with duplicated information
duplicated_tourneys <- c('1973-1098', '1970-9205', '1981-1099',
                         '1990-W-WT-ITA-01A-1990', '1991-W-WT-URS-01A-1991',
                         '1991-W-WT-USA-19A-1991', '1992-W-WT-USA-02A-1992')

tennis_results <- tennis_results %>%
  filter(!(str_detect(tourney_id, "-615") | str_detect(tourney_id, "-8888")),
         !(tourney_id %in% duplicated_tourneys)) %>%
  rbind(clean_1973_surbiton, clean_1981_johannesburg,
        clean_1990_taranto, clean_1991_stpetersburg,
        clean_1991_oakland, clean_1992_oklahoma)

# replacing mistaken entries
tennis_results[26765,13] = NA
tennis_results[26765,12] = 6
tennis_results[43756,12] = 9
# from looking at the original draw
# https://wtafiles.blob.core.windows.net/pdf/draws/archive/1983/702.pdf
# it can be deduced that they meant seed 9 instead of seed 96

# removing junior, challenger, exho results
# they are not the main pro tour
tennis_results <- tennis_results %>%
  filter(tourney_level != "J" & tourney_level != "CC" & tourney_level != "E")

# standardizing heights to cm
tennis_results <- tennis_results %>%
  mutate(winner_ht = ifelse(winner_ht < 100, winner_ht * 100, winner_ht),
         loser_ht = ifelse(loser_ht < 100, loser_ht * 100, loser_ht))

# adding a column of overall tournament winners for each match
winners <- tennis_results %>%
  filter(round == 'F') %>%
  mutate(tourney_winner = winner_name) %>%
  select(tour:month, tourney_winner)
tennis_results <- tennis_results %>%
  left_join(winners)
```

EDA

Number and Type of Tournaments

```
tennis_results %>%
  group_by(tour, year) %>%
  summarize(num_tourneys = n_distinct(tourney_name)) %>%
  ggplot(aes(x = year, y = num_tourneys, color = tour)) +
  geom_line()
```



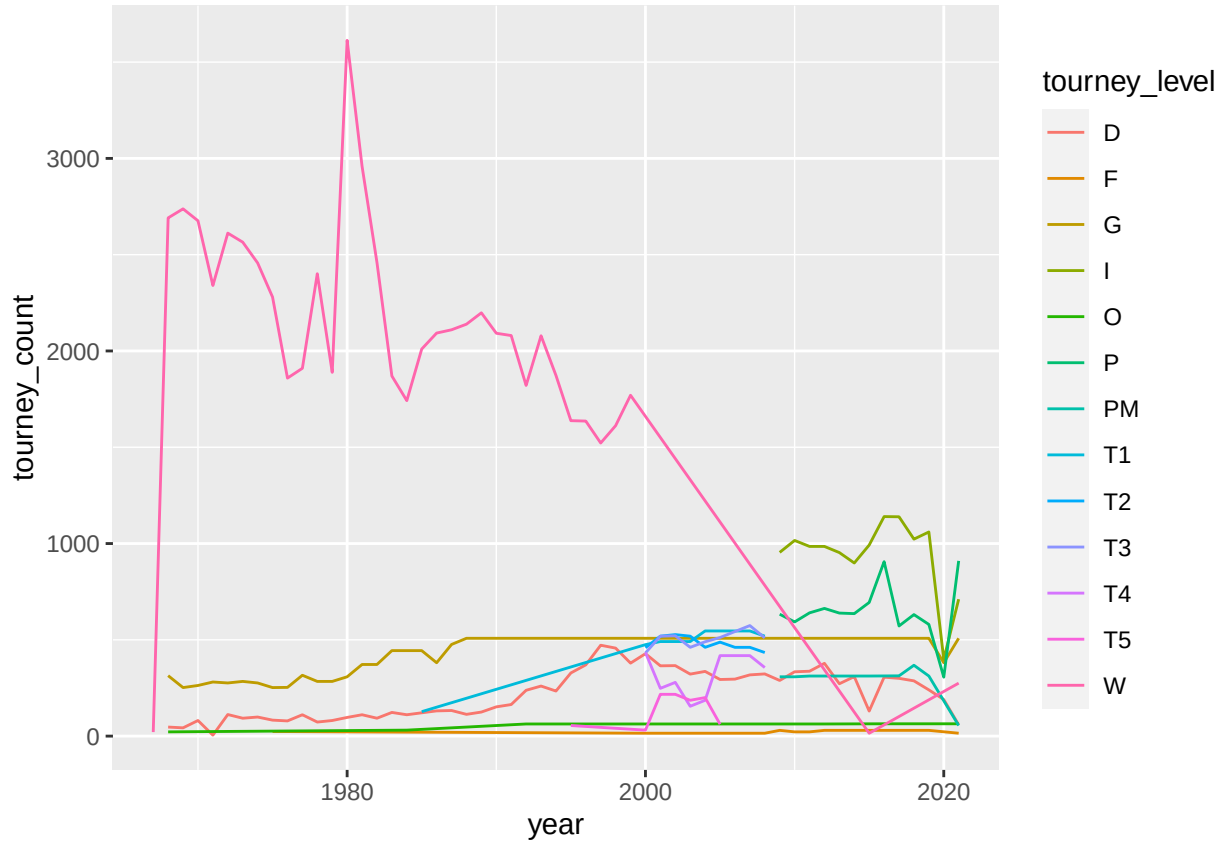

```
# there are 2 tournaments technically in 1967
# those are considered part of the 1968 WTA Tour

tennis_results %>%
  group_by(tour, year) %>%
  summarize(num_tourneys = n_distinct(tourney_name))
```

```
## # A tibble: 109 x 3
## # Groups:   tour [2]
##    tour   year num_tourneys
##    <chr> <dbl>        <int>
##  1 ATP    1968           67
##  2 ATP    1969           84
##  3 ATP    1970          113
```

```
##  4 ATP    1971          130
##  5 ATP    1972          142
##  6 ATP    1973          163
##  7 ATP    1974          166
##  8 ATP    1975          149
##  9 ATP    1976          152
## 10 ATP    1977          147
## # ... with 99 more rows
```

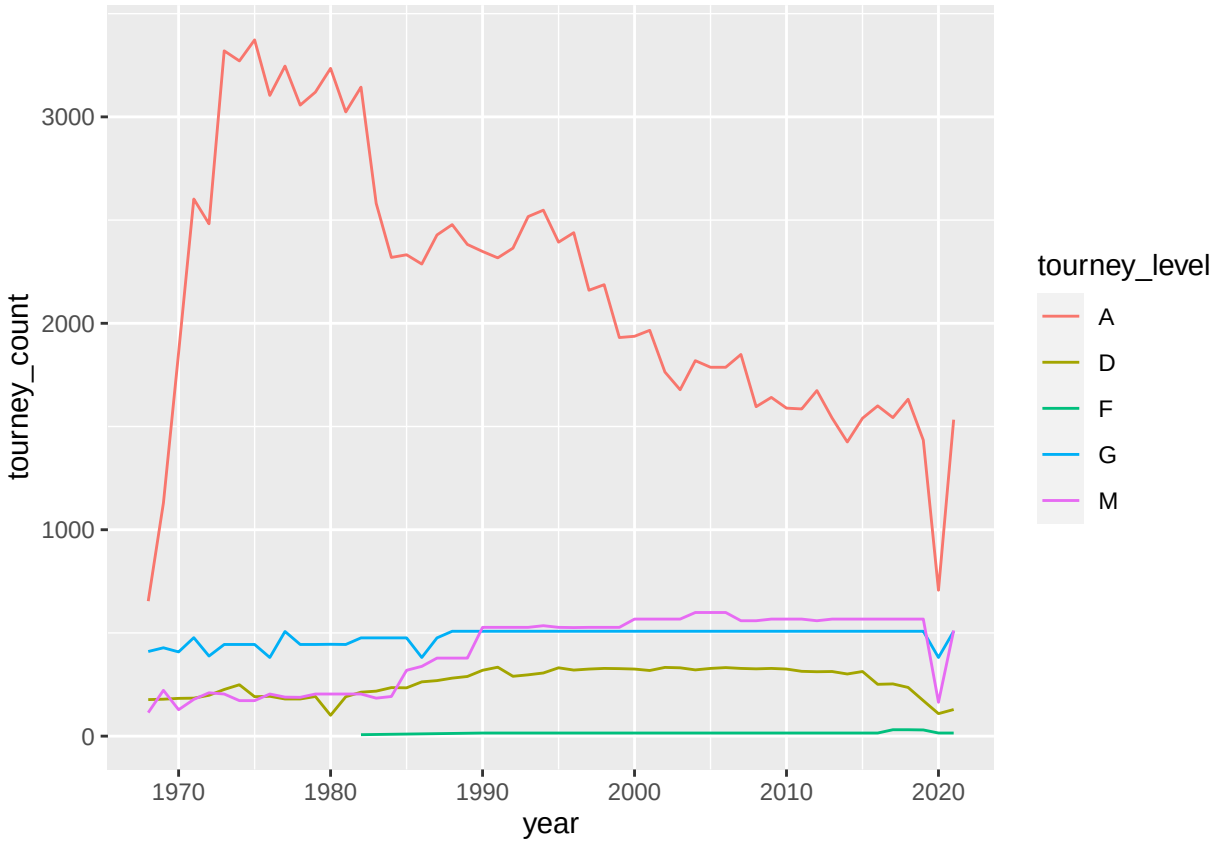
```
# the drop off in tournaments in 2020/2021 is explained by the pandemic
# why does WTA 2015 drop to 107?
# look at year by year data to see what tournaments are included in data
```

```
tennis_results %>%
  group_by(tour, year, tourney_level) %>%
  summarize(tourney_count = n()) %>%
  ungroup() %>%
  filter(tour == "WTA") %>%
  ggplot(aes(x = year, y = tourney_count, color = tourney_level)) +
  geom_line()
```



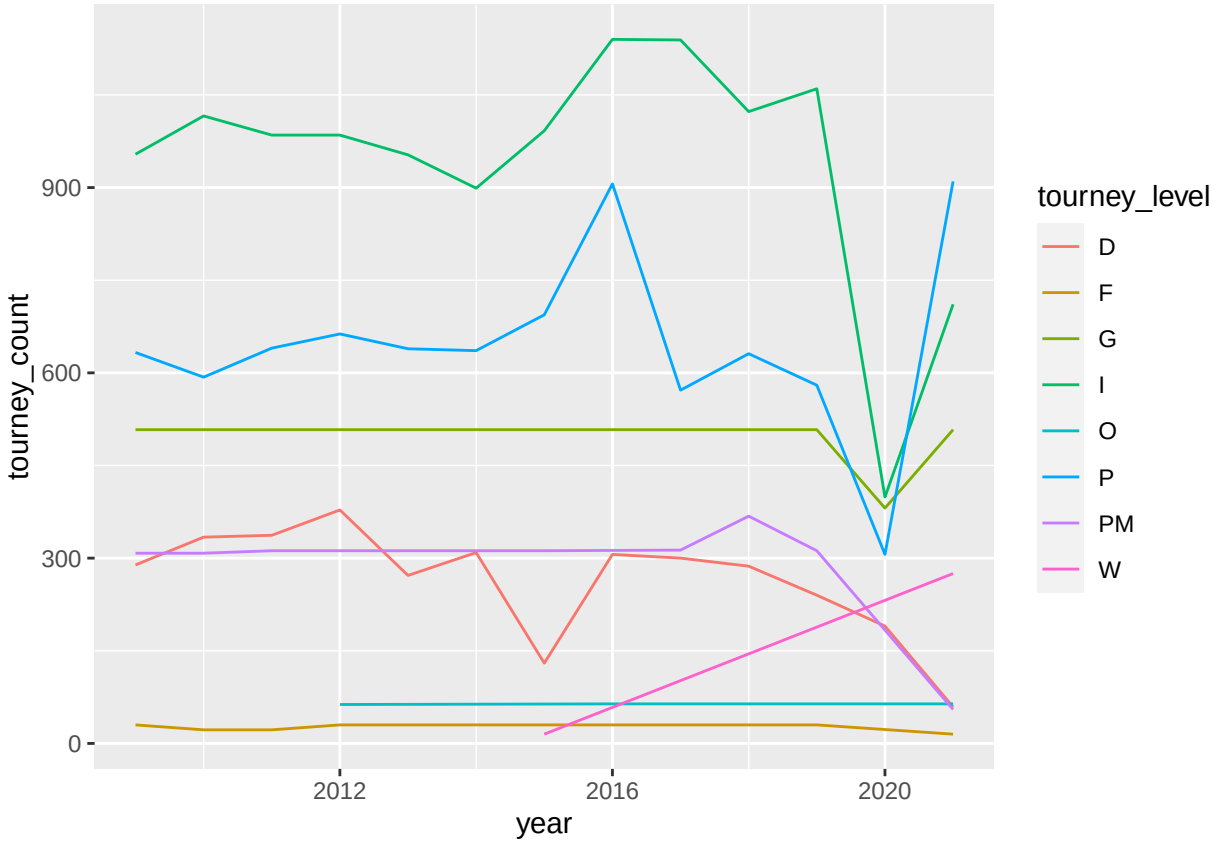

```
tennis_results %>%
  group_by(tour, year, tourney_level) %>%
  summarize(tourney_count = n()) %>%
  ungroup() %>%
```

```
  filter(tour == "ATP") %>%
  ggplot(aes(x = year, y = tourney_count, color = tourney_level)) +
  geom_line()
```

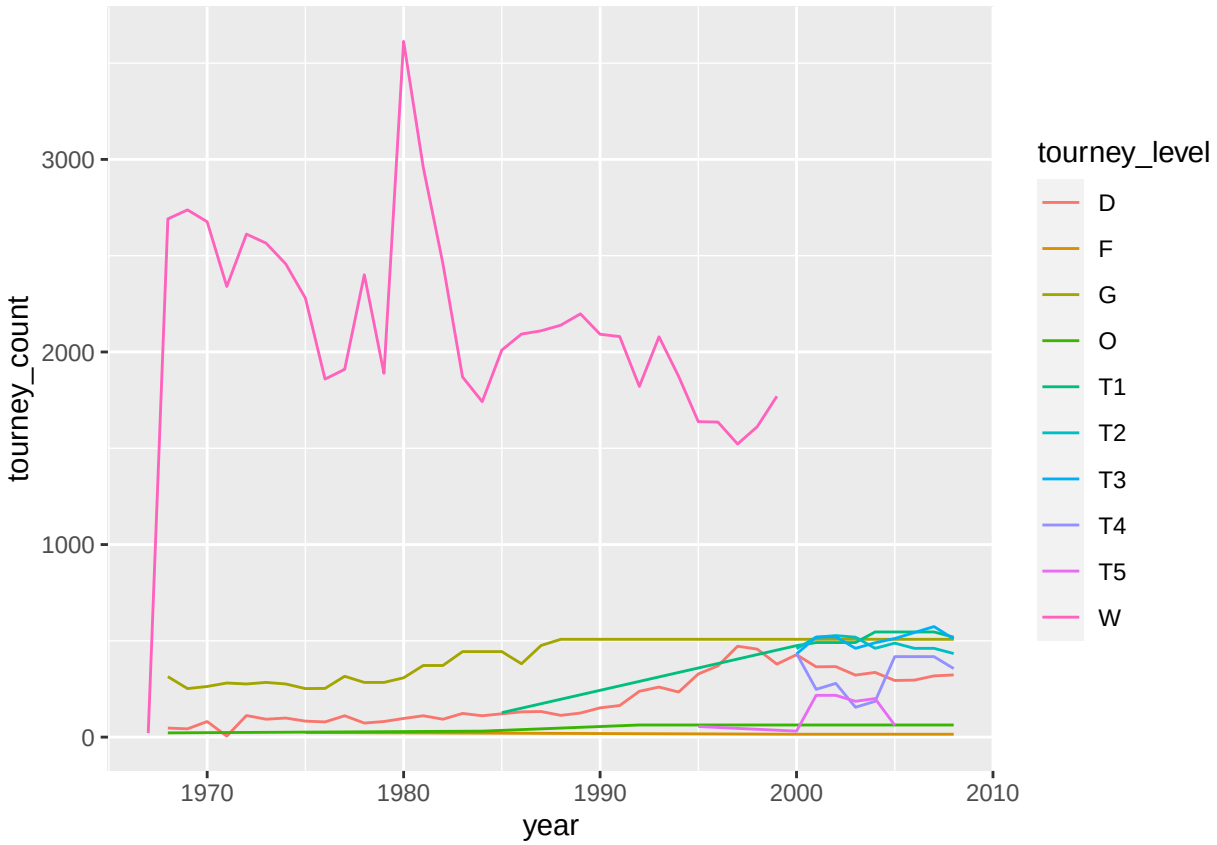



```
# what are the W and A tournaments?
# some of the connections are misleading
# no data points for some periods of time for some colors
# see readme for explanations of tourney_level
```

```
# in 2009, WTA started classifying tournaments differently
tennis_results %>%
  group_by(tour, year, tourney_level) %>%
  summarize(tourney_count = n()) %>%
  ungroup() %>%
  filter(tour == "WTA", year >= 2009) %>%
  ggplot(aes(x = year, y = tourney_count, color = tourney_level)) +
  geom_line()
```



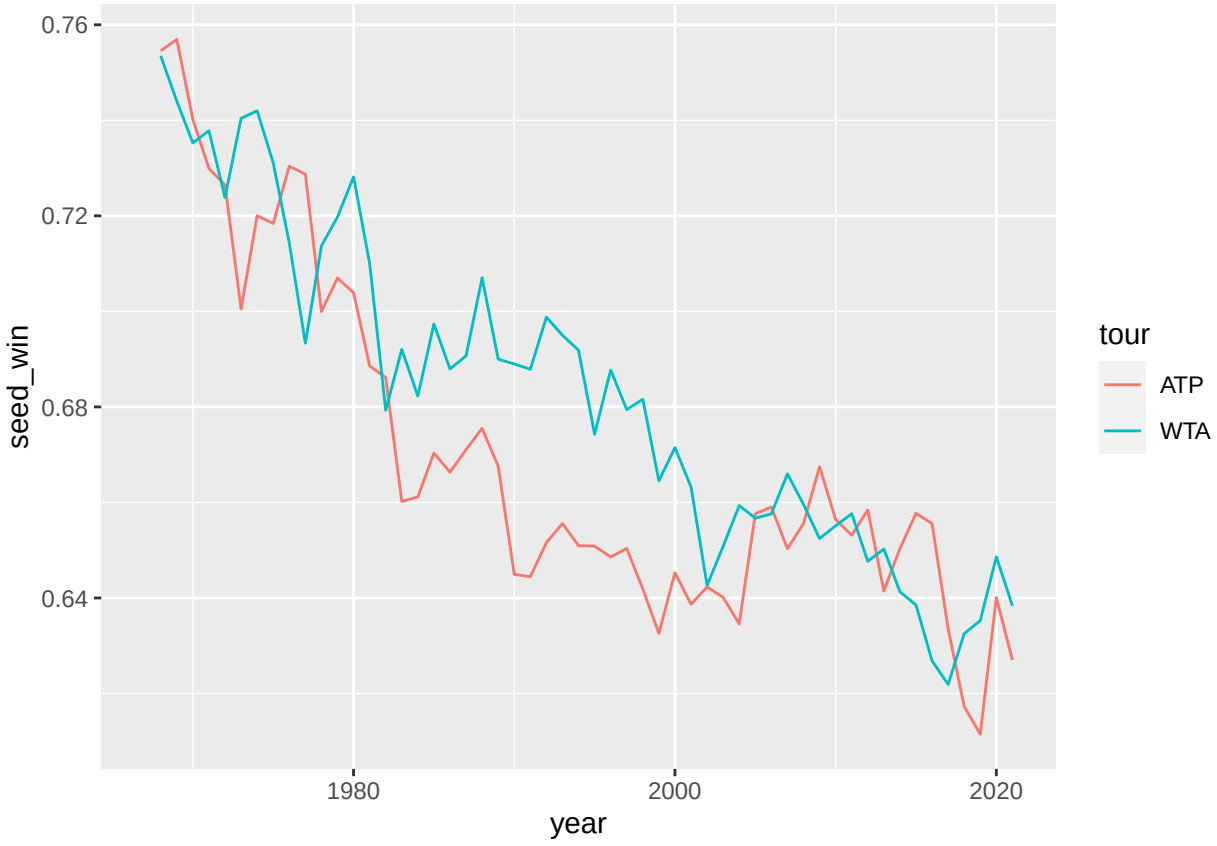

```
tennis_results %>%
  group_by(tour, year, tourney_level) %>%
  summarize(tourney_count = n()) %>%
  ungroup() %>%
  filter(tour == "WTA", year < 2009) %>%
  ggplot(aes(x = year, y = tourney_count, color = tourney_level)) +
  geom_line()
```




```
# why does W suddenly stop at 1999 and return in 2015
```

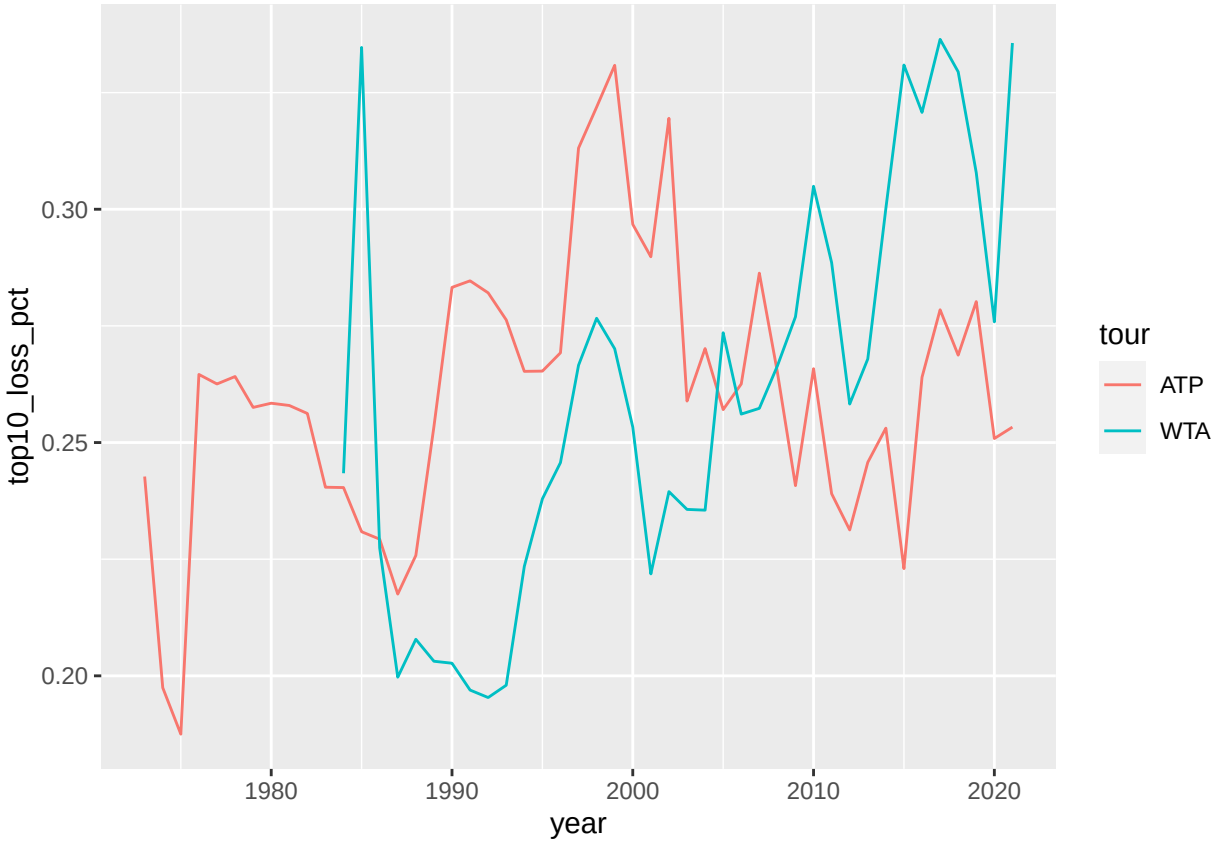
Measures of Tour (In)consistency

```
tennis_results %>%
  group_by(year, tour) %>%
  summarize(seed_win = sum(!is.na(winner_seed))/(sum(!is.na(winner_seed))+sum(!is.na(loser_seed)))) %>%
  ggplot(aes(x = year, y = seed_win, color = tour)) +
  geom_line()
```




```
# if anything, WTA has been more consistent!
```

```
tennis_results %>%
  group_by(year, tour) %>%
  filter(!is.na(loser_rank), !is.na(winner_rank)) %>%
  summarize(top10_loss_pct = sum(ifelse(loser_rank <= 10, 1, 0))/(sum(ifelse(loser_rank <= 10, 1, 0))+su
  ggplot(aes(x = year, y = top10_loss_pct, color = tour)) +
  geom_line()
```

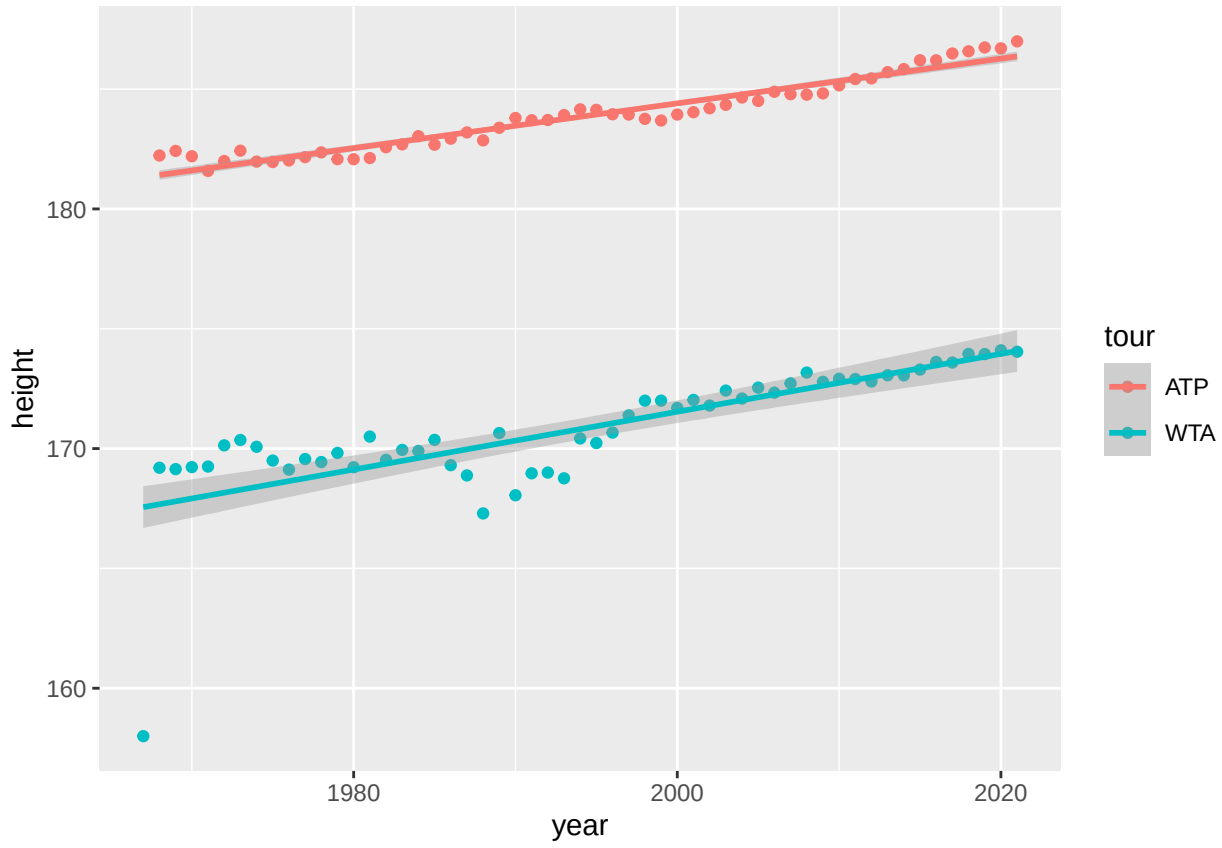



```
# in recent years, WTA top 10 less consistent but used to be more consistent
```

Tour Composition

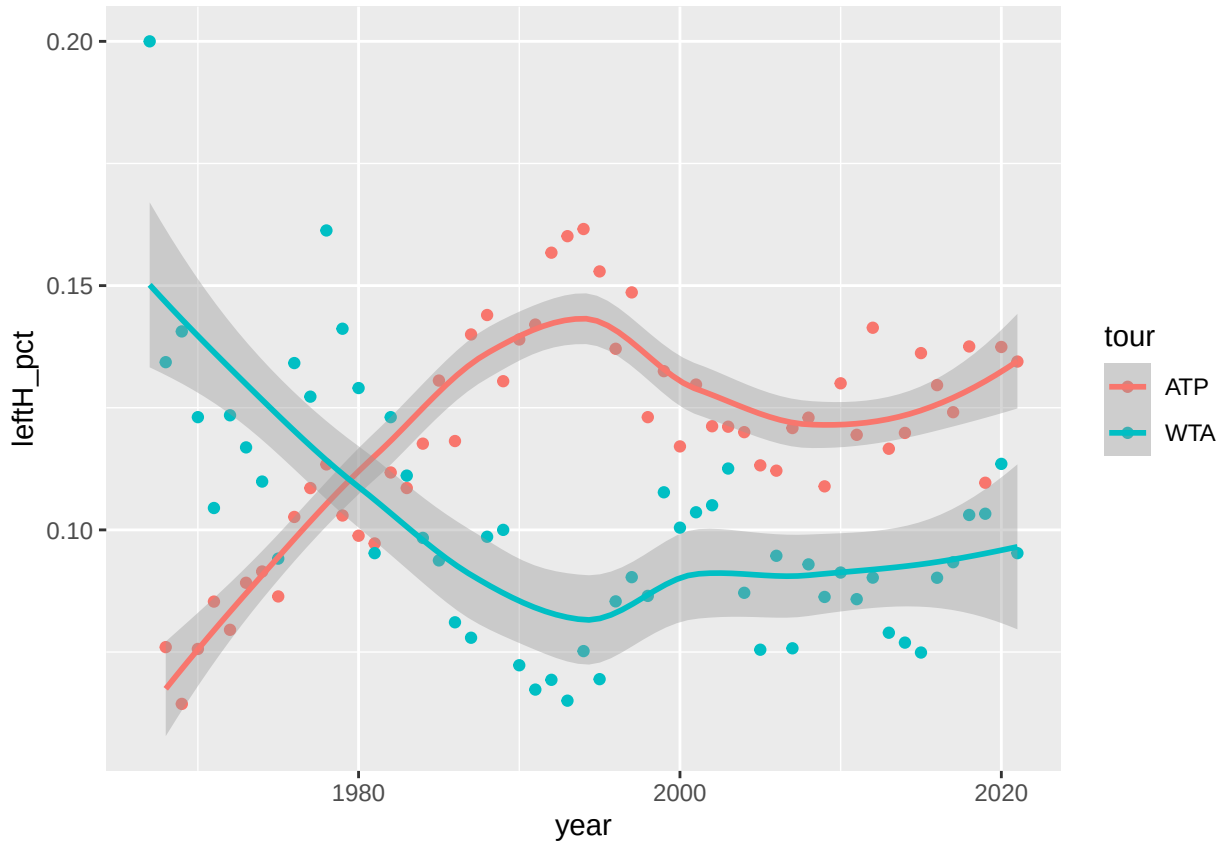
It would probably be better to use ranking data here than this tournament data, but we can get a sense of what's going on here.

```
tennis_results %>%
  group_by(tour, year) %>%
  distinct(winner_id, .keep_all = TRUE) %>%
  filter(!is.na(winner_ht)) %>%
  summarize(height = mean(winner_ht)) %>%
  ggplot(aes(x = year, y = height, color = tour)) +
  geom_point() +
  geom_smooth(method = "lm")
```



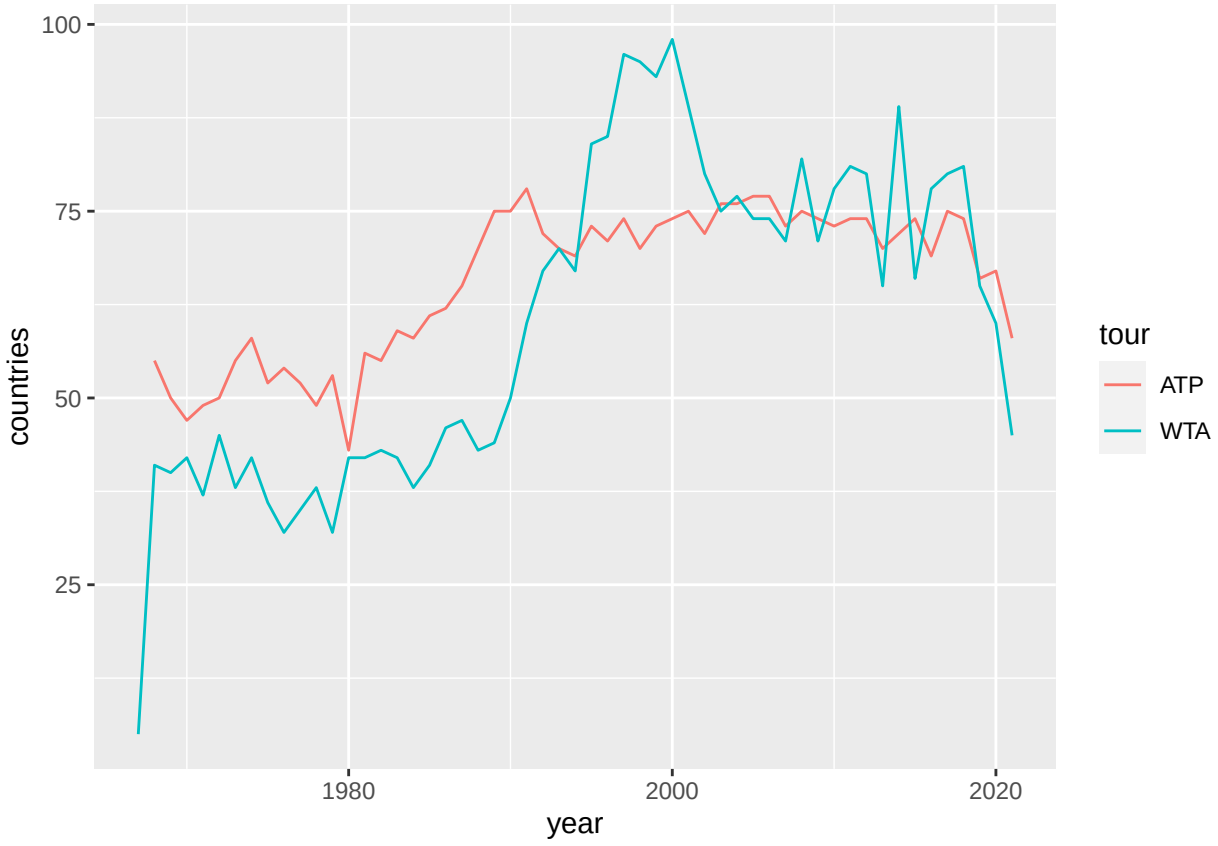

```
# height is increasing over time
```

```
tennis_results %>%
  group_by(year, tour) %>%
  distinct(winner_id, .keep_all = TRUE) %>%
  filter(winner_hand == "R" | winner_hand == "L") %>%
  summarize(leftH_pct = mean(winner_hand == "L")) %>%
  ggplot(aes(x = year, y = leftH_pct, color = tour)) +
  geom_point() +
  geom_smooth()
```



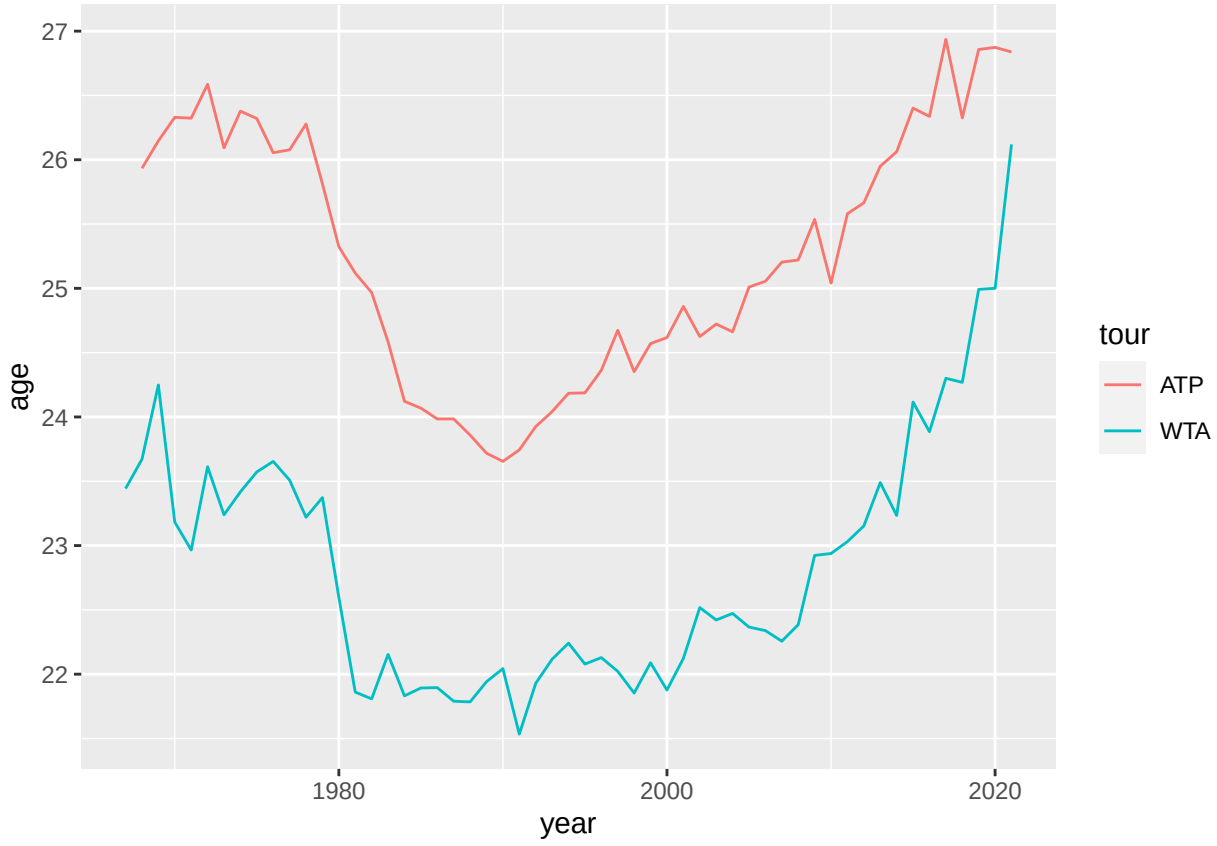

```
# lefties increasing in ATP and decreasing in WTA?
```

```
tennis_results %>%
  group_by(tour, year) %>%
  summarize(countries = n_distinct(winner_ioc)) %>%
  ggplot(aes(x = year, y = countries, color = tour)) +
  geom_line()
```



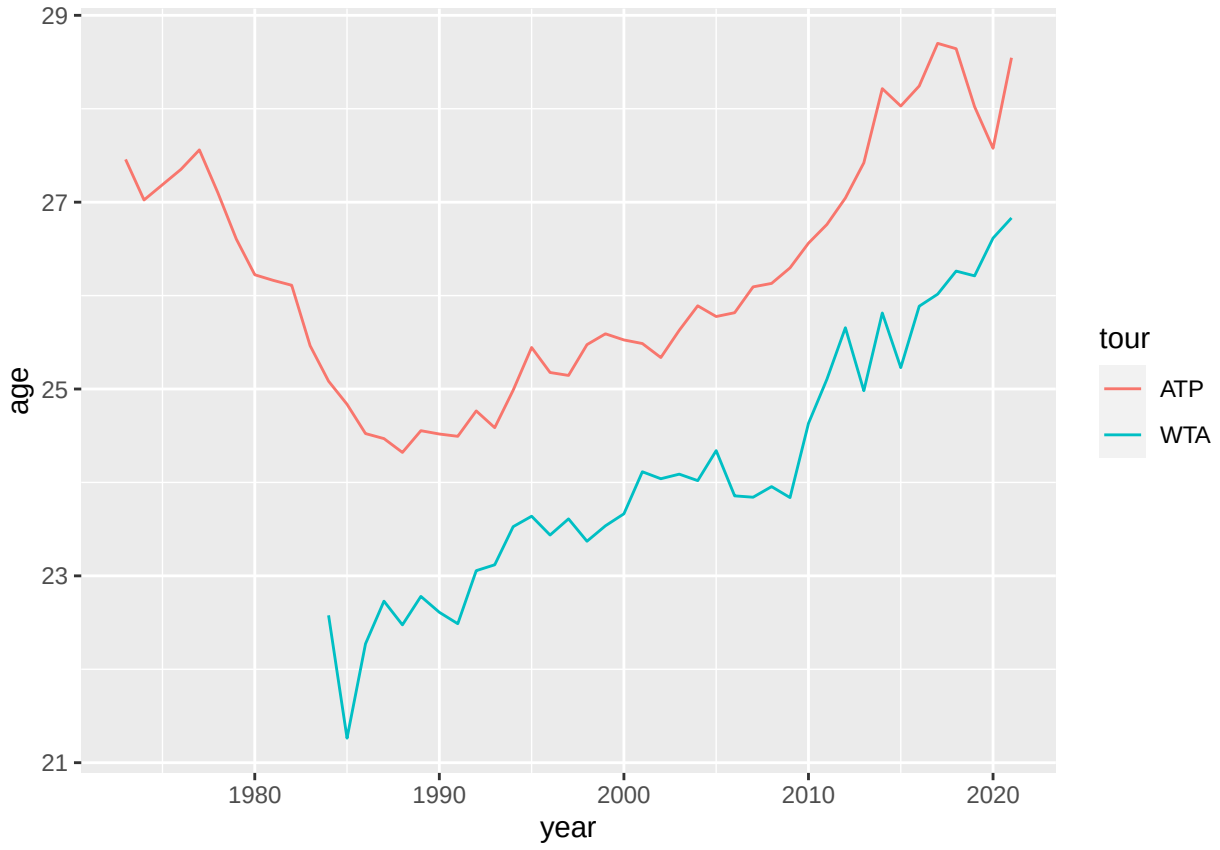

```
# increase in countries until around 1995, plateauing since then
```

```
tennis_results %>%
  group_by(tour, year) %>%
  distinct(winner_id, .keep_all = TRUE) %>%
  filter(!is.na(winner_age)) %>%
  summarize(age = mean(winner_age), count = n()) %>%
  ggplot(aes(x = year, y = age, color = tour)) +
  geom_line()
```



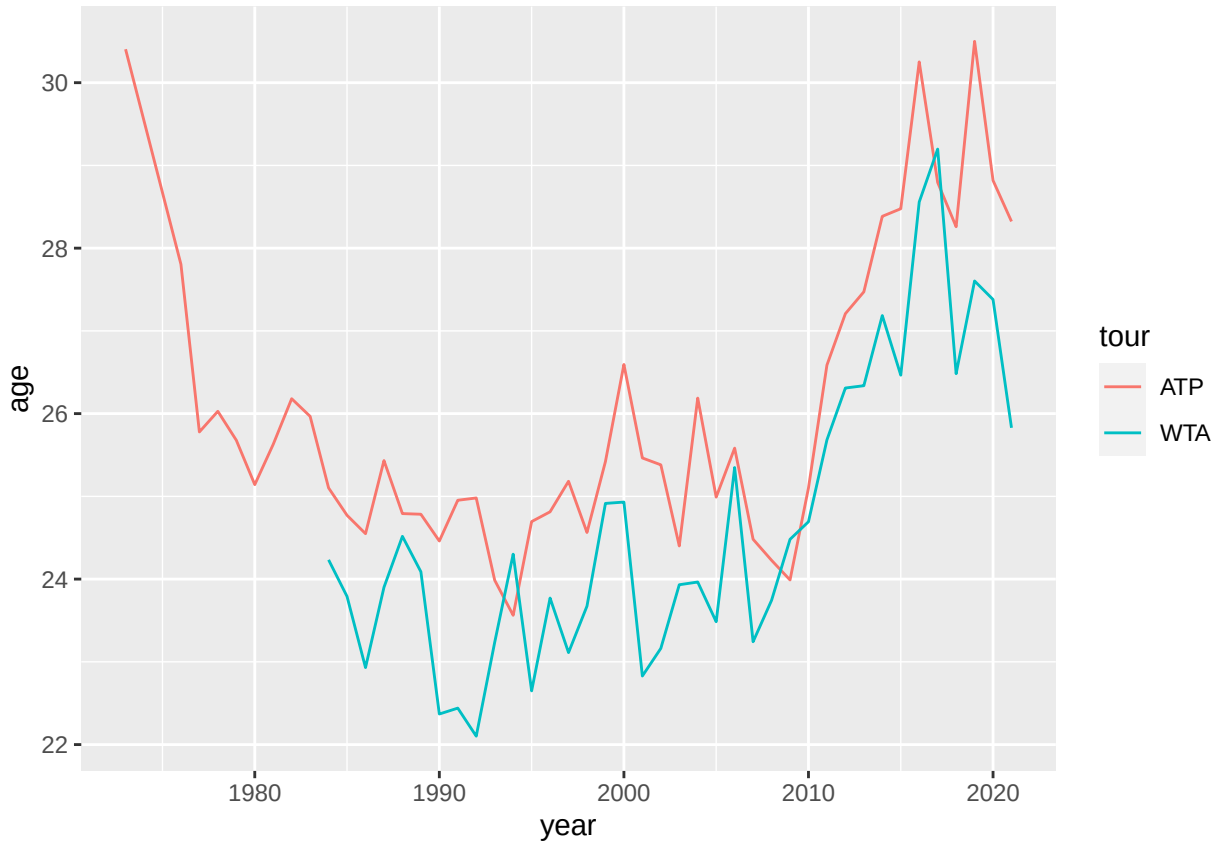

```
# since 1990, getting older consistently
# not sure why tour was younger before 1990
```

```
tennis_results %>%
  group_by(tour, year) %>%
  distinct(winner_id, .keep_all = TRUE) %>%
  filter(!is.na(winner_age), winner_rank <= 100) %>%
  summarize(age = mean(winner_age), count = n()) %>%
  ggplot(aes(x = year, y = age, color = tour)) +
  geom_line()
```




```
# why no data for WTA before 1984?
# top 100 also getting older from 1990 onward
```

```
tennis_results %>%
  group_by(tour, year) %>%
  distinct(winner_id, .keep_all = TRUE) %>%
  filter(!is.na(winner_age), winner_rank <= 10) %>%
  summarize(age = mean(winner_age), count = n()) %>%
  ggplot(aes(x = year, y = age, color = tour)) +
  geom_line()
```




```
# top 10 is more variable with smaller sample
# general trend of getting older still holds
# 1984 cut off again for WTA
```

Playing Style

Look at this **ATP cluster analysis**. So cool!!! Exactly what we learned/are learning in class in practice. We could do something similar for WTA and compare the results. (Is that something valid to do?)