rankings

Group E

12/12/2021

```
knitr::opts_chunk$set(message = FALSE, warning = FALSE)
```

```
library(readr)
rank_URLs <- c("https://raw.githubusercontent.com/JeffSackmann/tennis_wta/master/wta_rankings_90s.csv",
               "https://raw.githubusercontent.com/JeffSackmann/tennis_wta/master/wta_rankings_00s.csv",
               "https://raw.githubusercontent.com/JeffSackmann/tennis_wta/master/wta_rankings_10s.csv",
               "https://raw.githubusercontent.com/JeffSackmann/tennis_wta/master/wta_rankings_20s.csv",
               "https://raw.githubusercontent.com/JeffSackmann/tennis_wta/master/wta_rankings_current.cs
wtarank <- read_csv(rank_URLs)
```

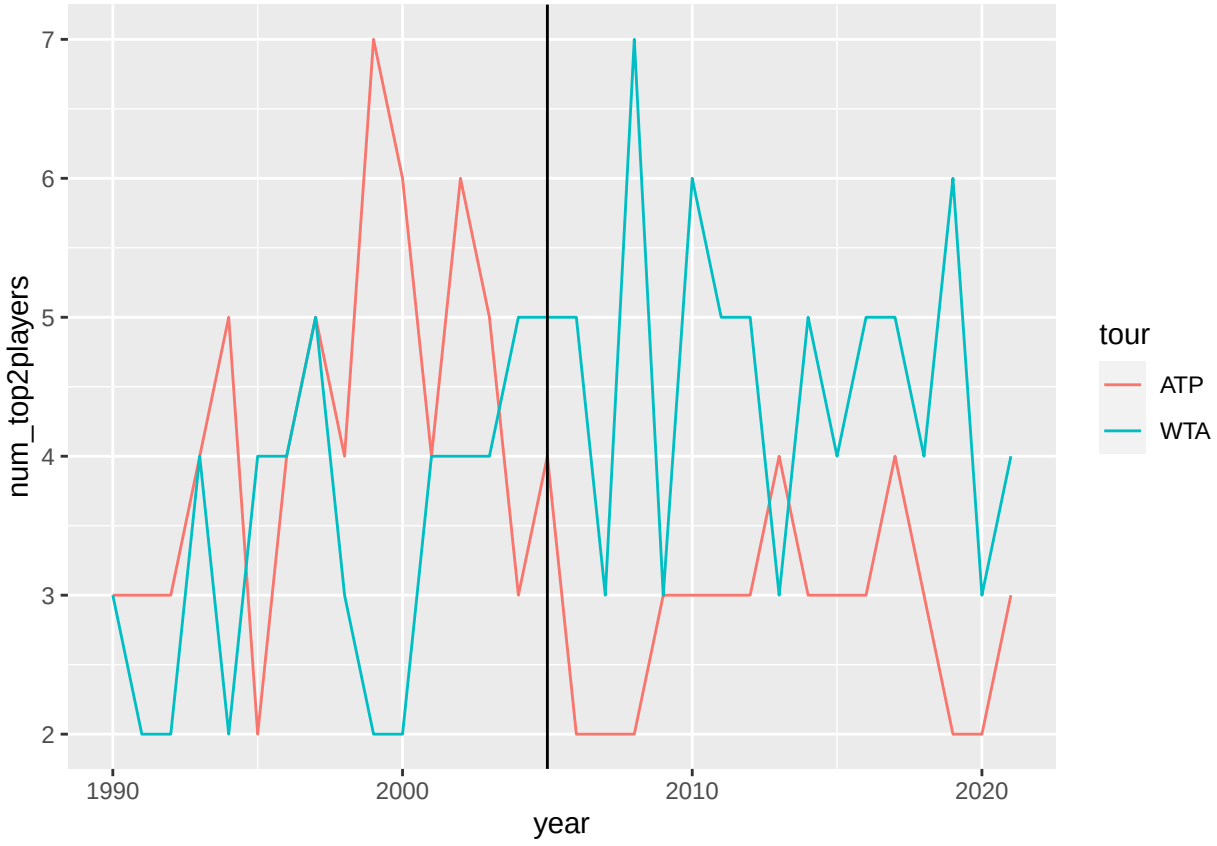
```
# 20s file is missing first row with variable names so loaded separately
atp_rankings_90s <- read_csv("https://raw.githubusercontent.com/JeffSackmann/tennis_atp/master/atp_ranki
atp_rankings_00s <- read_csv("https://raw.githubusercontent.com/JeffSackmann/tennis_atp/master/atp_ranki
atp_rankings_10s <- read_csv("https://raw.githubusercontent.com/JeffSackmann/tennis_atp/master/atp_ranki
atp_rankings_20s <- read_csv("https://raw.githubusercontent.com/JeffSackmann/tennis_atp/master/atp_ranki
    col_names = FALSE)
colnames(atp_rankings_20s) <- c("ranking_date", "rank", "player", "points")
atp_rankings_current <- read_csv("https://raw.githubusercontent.com/JeffSackmann/tennis_atp/master/atp_r
atprank <- rbind(atp_rankings_90s,
                 atp_rankings_00s,
                 atp_rankings_10s,
                 atp_rankings_20s,
                 atp_rankings_current)
```

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

atprank <- atprank %>%
  mutate(tours = NA, tour = "ATP")

# moving tour column to front for ease
wtarank <- wtarank[,c(6,5,1:4)]
atprank <- atprank[,c(5,6,1:4)]

# combining the datasets
tennis_rankings <- rbind(wtarank, atprank)

# making date objects from date
library(lubridate)
```

```r
tennis_rankings <- tennis_rankings %>%
  mutate(ranking_date = ymd(ranking_date)) %>%
  mutate(year = year(ranking_date)) %>%
  mutate(month = month(ranking_date)) %>%
  mutate(week = week(ranking_date))

# reorganizing date columns together
tennis_rankings <- tennis_rankings[,c(1:3,7:9,4:6)]
```

```r
library(gganimate)
library(ggimage)
library(gifski)
ranking_dominance_animate <- tennis_rankings %>%
  filter(rank <= 100, week == 52, year >= 1996) %>%
  mutate(top10 = ifelse(rank <= 10, 1, 0)) %>%
  group_by(top10, year, tour) %>%
  summarize(pts = sum(points, na.rm = TRUE)) %>%
  group_by(tour, year) %>%
  summarize(ranking_dominance = pts[2]/sum(pts)) %>%
  mutate(image = tour) %>%
  mutate(image = case_when(
    image == "WTA" ~ "https://upload.wikimedia.org/wikipedia/en/thumb/5/5f/Women%27s_Tennis_Association_
    image == "ATP" ~ "https://upload.wikimedia.org/wikipedia/en/thumb/3/3f/ATP_Tour_logo.svg/800px-ATP_T
  )) %>%
  ggplot(aes(x = year, y = ranking_dominance)) +
  geom_line(aes(color = tour)) +
  geom_image(aes(image = image), size = 0.1) +
  guides(col = FALSE) +
  transition_reveal(year)
anim_save("ranking_dominance_animate.gif", animation = ranking_dominance_animate, path = "~/Comp Stats/C
```

```r
tennis_rankings %>%
  group_by(tour, year) %>%
  filter(rank <= 2) %>%
  summarize(num_top2players = n_distinct(player)) %>%
  ggplot(aes(x = year, y = num_top2players)) +
  geom_line(aes(color = tour)) +
  geom_vline(xintercept = 2005)
```




```
tennis_rankings %>%
  group_by(tour, year) %>%
  filter(rank <= 2, year <= 2005) %>%
  summarize(num_top2players = n_distinct(player)) %>%
  summarize(mean_top2players = mean(num_top2players))
```

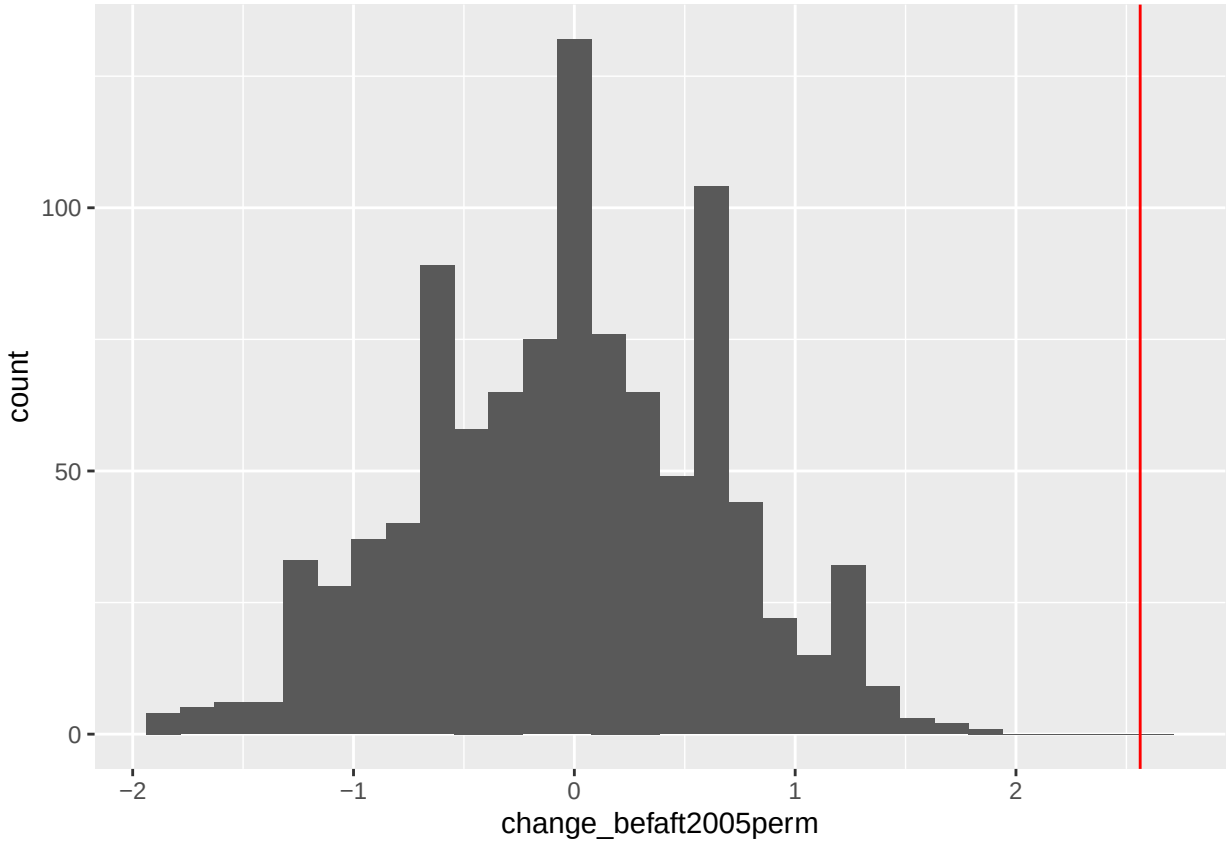
```
## # A tibble: 2 x 2
##   tour  mean_top2players
##   <chr>            <dbl>
## 1 ATP               4.25
## 2 WTA               3.44
```

```
tennis_rankings %>%
  group_by(tour, year) %>%
  filter(rank <= 2, year > 2005) %>%
  summarize(num_top2players = n_distinct(player)) %>%
  summarize(mean_top2players = mean(num_top2players))
```

```
## # A tibble: 2 x 2
##   tour  mean_top2players
##   <chr>            <dbl>
## 1 ATP               2.81
## 2 WTA               4.56
```

```r
top2byyear <- tennis_rankings %>%
  group_by(tour, year) %>%
  filter(rank <= 2)

top2_rank_func <- function(.x){
  top2byyear %>%
    summarize(num_top2players = n_distinct(player)) %>%
    mutate(bef_2005 = ifelse(year <= 2005, 1, 0)) %>%
    ungroup() %>%
    group_by(bef_2005) %>%
    mutate(num_top2perm = sample(num_top2players, replace = FALSE)) %>%
    group_by(tour, bef_2005) %>%
    summarize(avg_bef2005 = mean(num_top2players),
              avg_bef2005_perm = mean(num_top2perm)) %>%
    group_by(bef_2005) %>%
    summarize(diff_bef2005 = diff(avg_bef2005),
              diff_bef2005perm = diff(avg_bef2005_perm)) %>%
    summarize(change_befaft2005 = diff_bef2005[1]-diff_bef2005[2],
              change_befaft2005perm = diff_bef2005perm[1]-diff_bef2005perm[2])
}
```

```r
set.seed(47)
perm_diff_top2 <- map_df(1:1000, top2_rank_func)

perm_diff_top2 %>%
  ggplot() +
  geom_histogram(aes(x = change_befaft2005perm)) +
  geom_vline(aes(xintercept = change_befaft2005), color = "red")
```




```
perm_diff_top2 %>%
  summarize(pval = 1-sum(abs(change_befaft2005) > change_befaft2005perm) / 1000)
```

```
## # A tibble: 1 x 1
##    pval
##   <dbl>
## 1     0
```