Ultra Running: 2021/10/26

Contents

1 Loading in the Data

```
library(tidyverse)
library(tidytuesdayR)
library(lubridate)

f.ultra = paste0(
  'https://raw.githubusercontent.com/rfordatascience/tidytuesday',
  '/master/data/2021/2021-10-26/ultra_rankings.csv'
)
ultra_rankings = readr::read_csv(f.ultra)

f.race = paste0(
  'https://raw.githubusercontent.com/rfordatascience/tidytuesday',
  '/master/data/2021/2021-10-26/race.csv'
)
race = readr::read_csv(f.race)

full_set = ultra_rankings %>%
  inner_join(race)
```

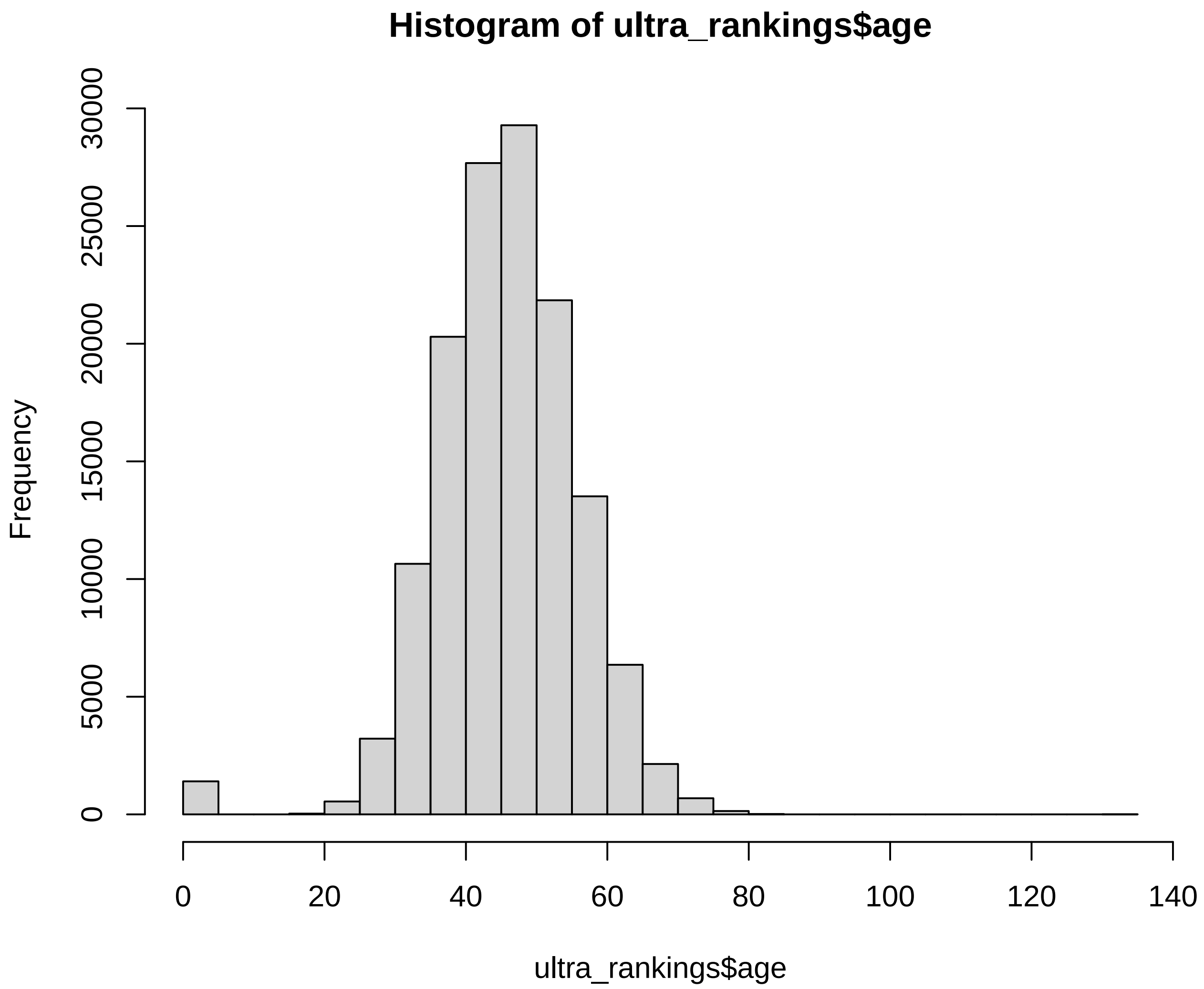
2 Data Manipulation and Wrangling

```
ultra_rankings = ultra_rankings %>%
  mutate(race_year_id = as.factor(race_year_id))

top_150_races = ultra_rankings %>%
  group_by(race_year_id) %>%
  summarise(n = n()) %>%
  filter(n >= 1633)
```

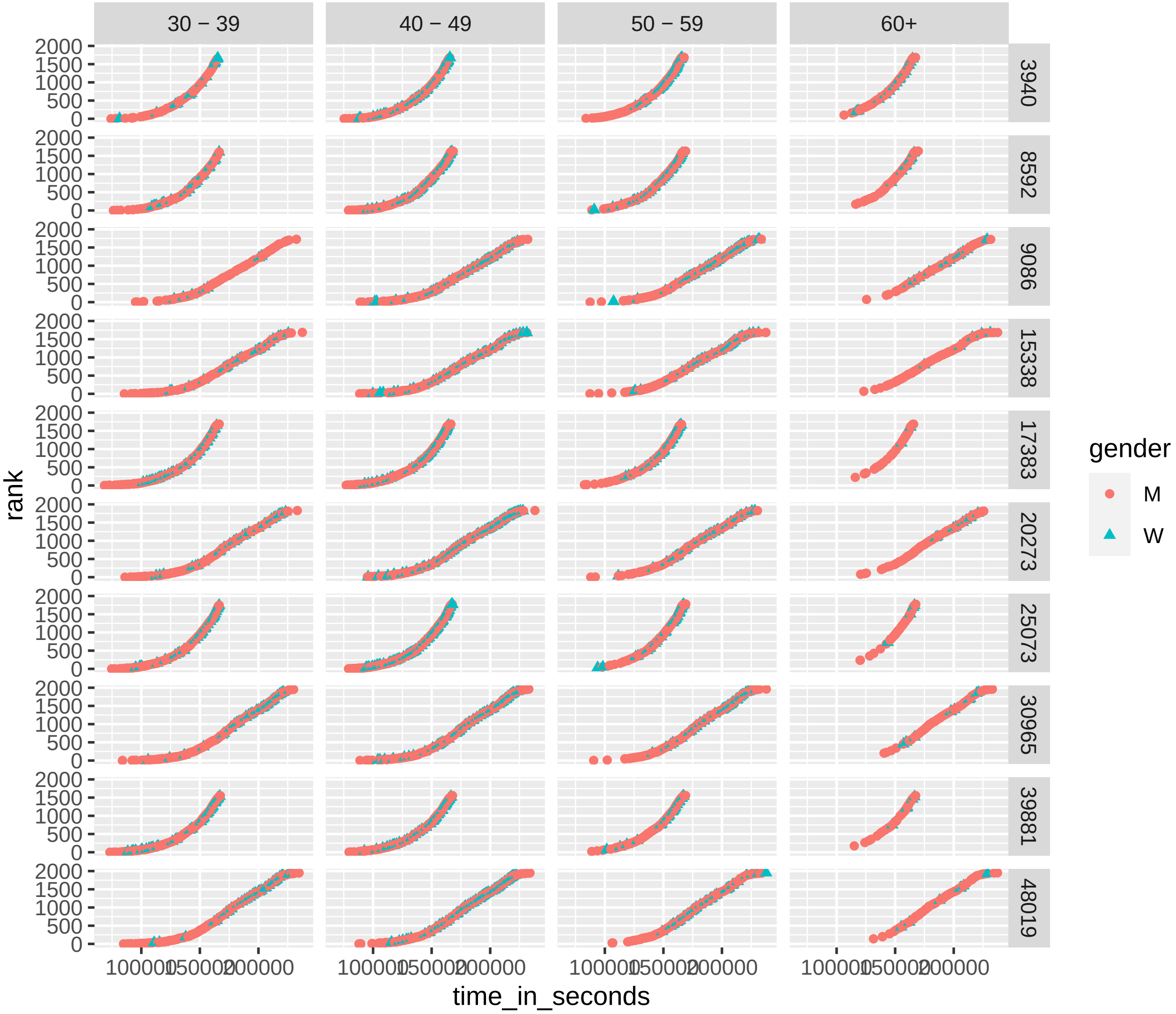
3 Analysis I: Influence of Age Group on Race Performance by Gender

```
hist(ultra_rankings$age)
```



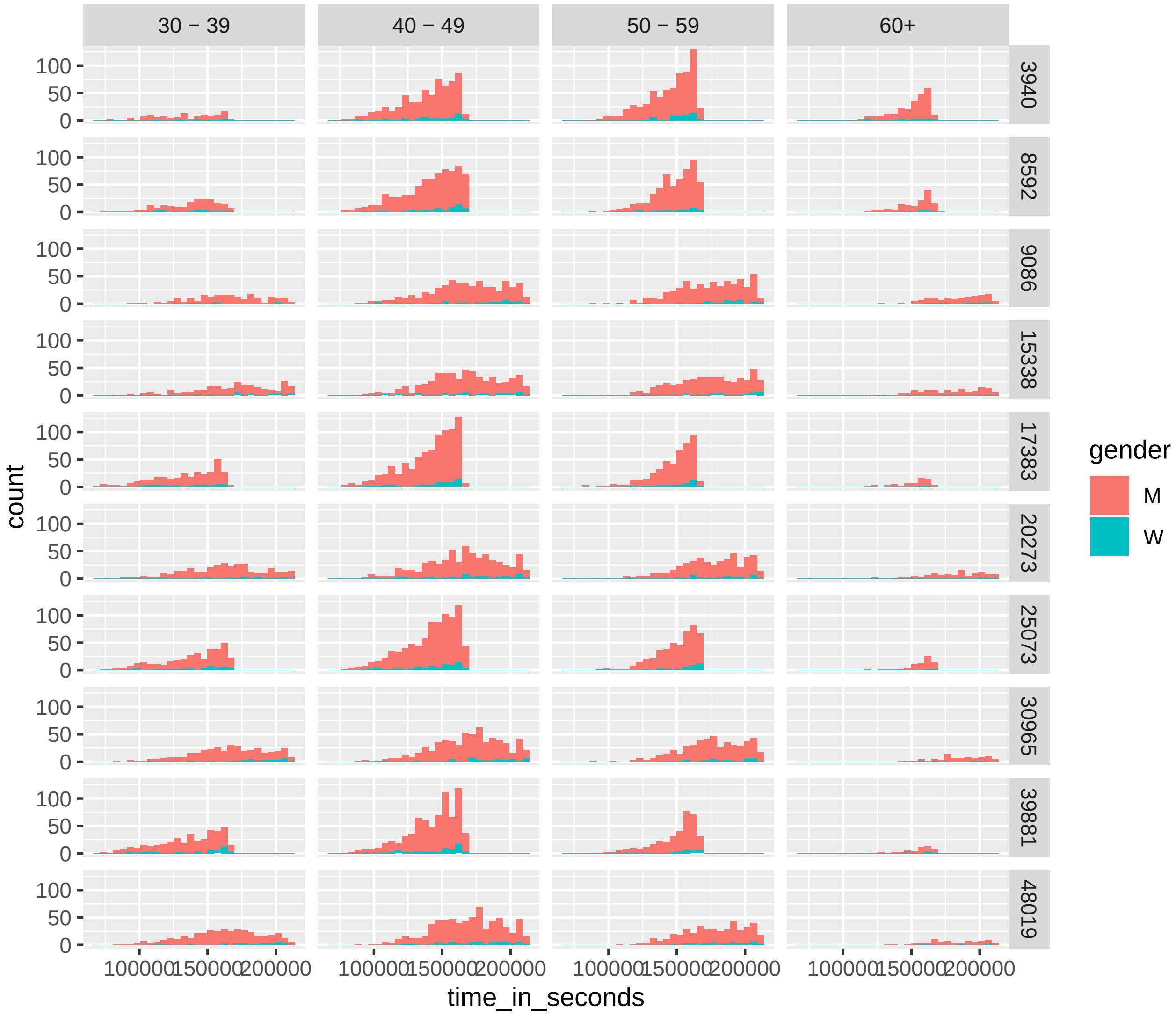
```
ultra_rankings %>%
  mutate(age_group = ifelse(age >= 30 & age < 40, '30 - 39',
                     ifelse(age >= 40 & age < 50, '40 - 49',
                     ifelse(age >= 50 & age < 60, '50 - 59',
                     ifelse(age >= 60, '60+', 'undefined'))))) %>%
  select(-c(runner, time, nationality, age)) %>%
  filter(age_group != 'undefined') %>%
  filter(race_year_id %in% top_150_races$race_year_id) %>%
  filter(gender %in% c('M', 'W')) %>%
  ggplot() +
    geom_point(aes(x=time_in_seconds, y=rank, colour=gender, shape=gender)) +
    facet_grid(race_year_id ~ age_group)
```

```
## Warning: Removed 4723 rows containing missing values (geom_point).
```



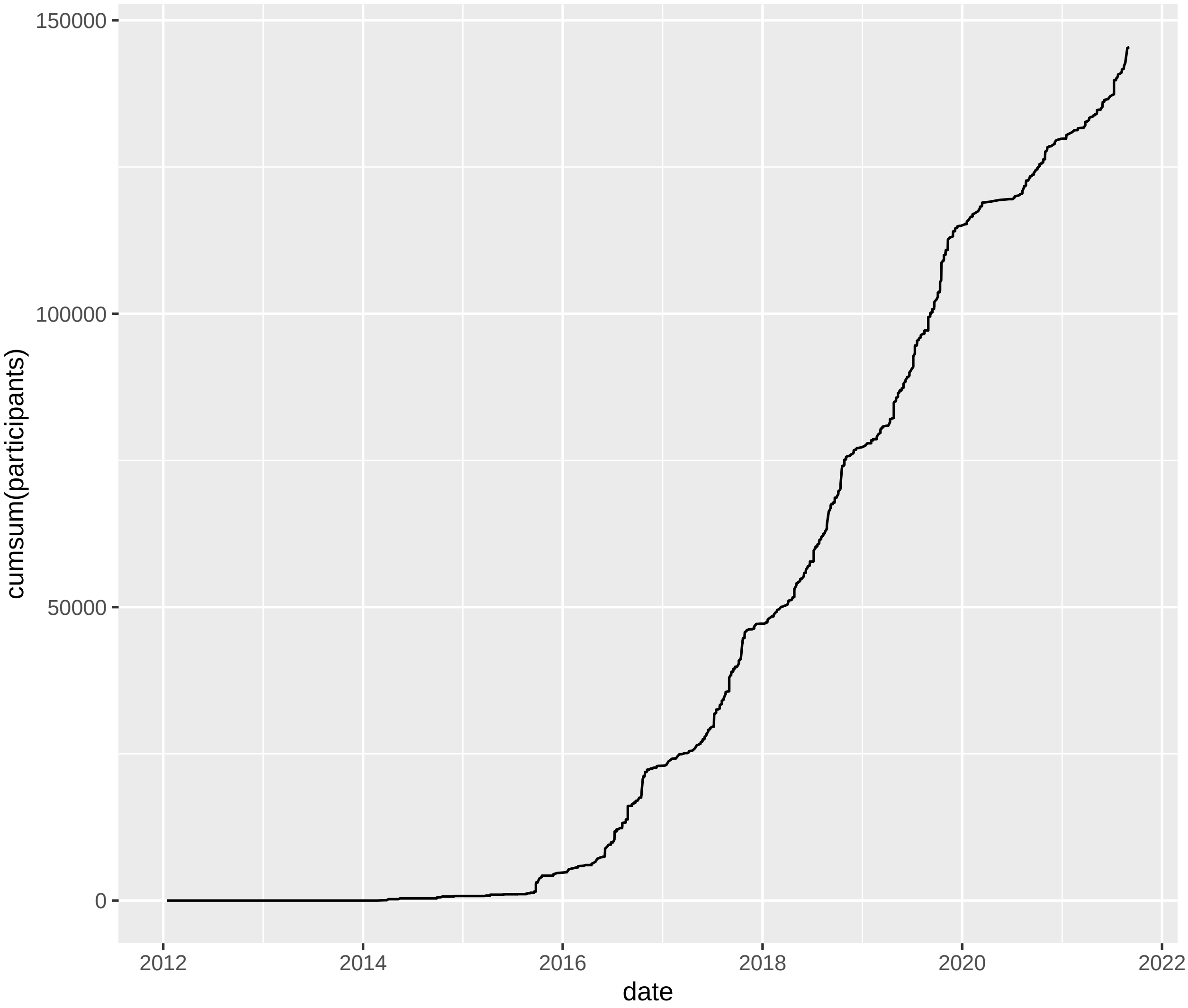
```
ultra_rankings %>%
  mutate(age_group = ifelse(age >= 30 & age < 40, '30 - 39',
                     ifelse(age >= 40 & age < 50, '40 - 49',
                     ifelse(age >= 50 & age < 60, '50 - 59',
                     ifelse(age >= 60, '60+', 'undefined'))))) %>%
  select(-c(runner, time, nationality, age)) %>%
  filter(time_in_seconds > quantile(time_in_seconds, 0.025, na.rm=TRUE),
         time_in_seconds < quantile(time_in_seconds, 0.975, na.rm=TRUE)) %>%
  filter(age_group != 'undefined') %>%
  filter(race_year_id %in% top_150_races$race_year_id) %>%
  filter(gender %in% c('M', 'W')) %>%
  ggplot() +
    geom_histogram(aes(x=time_in_seconds, fill=gender)) +
    facet_grid(race_year_id ~ age_group)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



4 Participation Trend over Time

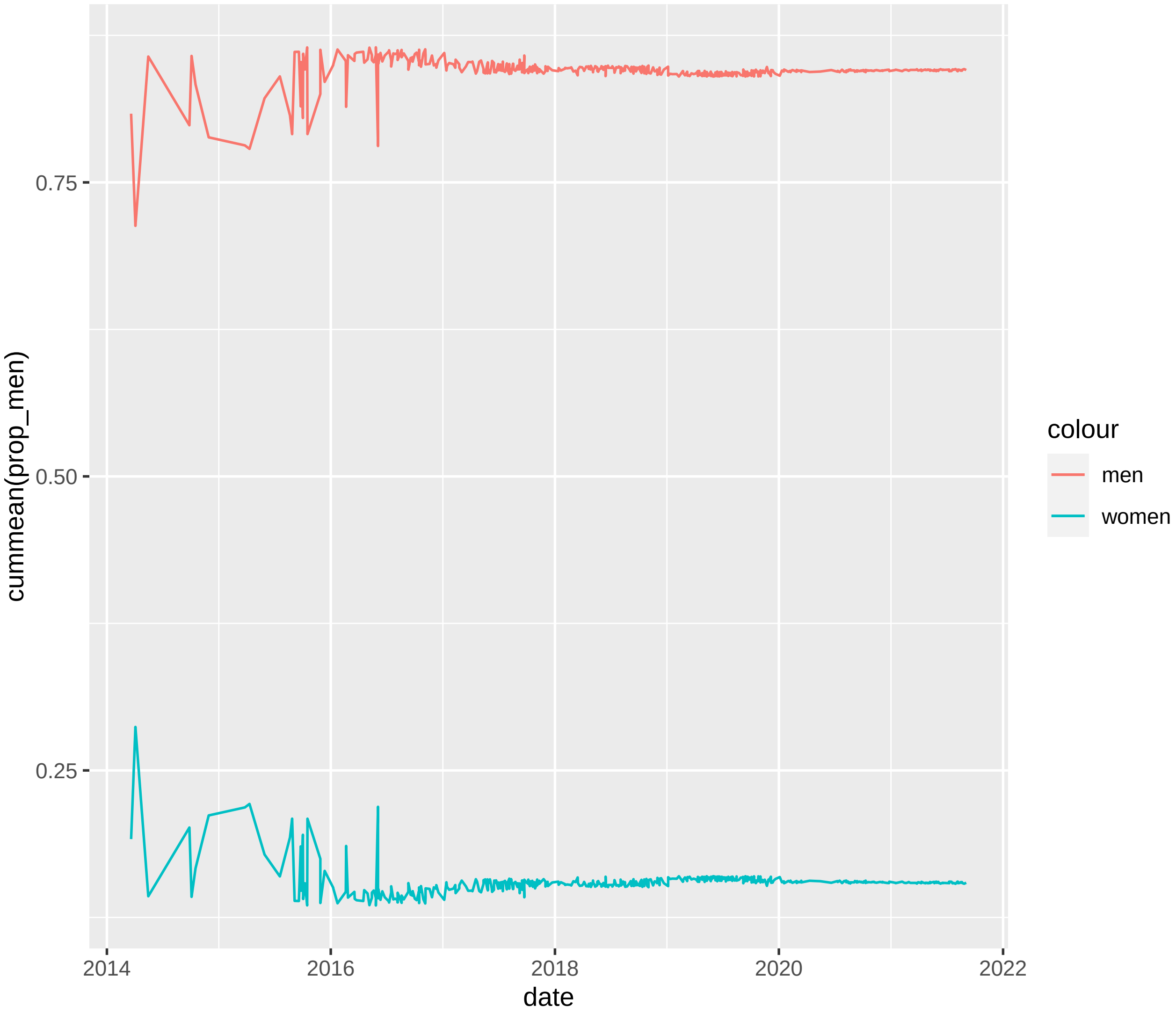
```
race %>%
  arrange(date) %>%
  ggplot() +
    geom_line(aes(x=date, y=cumsum(participants)))
```



One possibility is that less women are competing, hence explaining that the observed increase in womens performances could be due to the luck of a few speedy ladies. However, as we see over time, even though participation has increased, this proportion remains the same.

```
full_set %>%
  filter(gender %in% c('M', 'W')) %>%
  group_by(race_year_id) %>%
  summarise(prop_men = mean(gender == 'M'),
            prop_women = mean(gender == 'W')) %>%
  select(race_year_id, prop_men, prop_women) %>%
  right_join(race) %>%
  filter(participants > 0) %>%
  select(-c(event, race, city, country, start_time, participation, elevation_gain, elevation_lo
  ggplot(aes(x=date)) +
    geom_line(aes(y=cummean(prop_men), colour='men')) +
    geom_line(aes(y=cummean(prop_women), colour='women'))
```

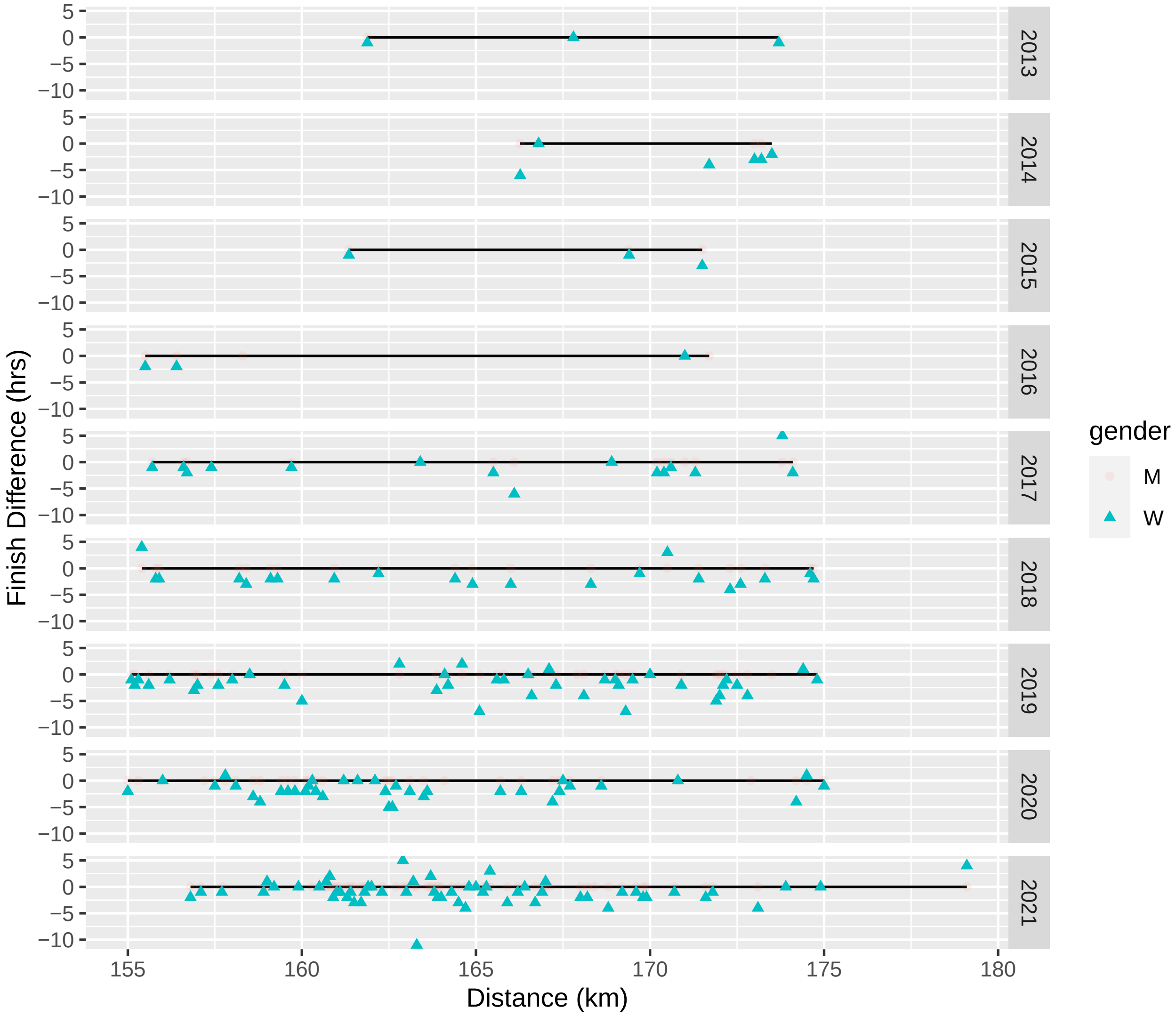
```
## Joining, by = "race_year_id"
```



Women have consistently outperformed their male counterparts in long-distance running, and the results become clearer as more women join year after year.

```
full_set %>%
  filter(gender %in% c('M', 'W') & distance > 150) %>%
  group_by(distance, gender) %>%
  summarise(mean_time = mean(time_in_seconds, na.rm=TRUE),
            date = max(date)) %>%
  mutate(gender_diff = (mean_time[gender == 'M'] - mean_time) %/% 3600,
         year = substr(lubridate::ymd(date), 1, 4)) %>%
  ggplot() +
    geom_line(aes(x=distance, y=0)) +
    geom_point(aes(x=distance, y=gender_diff, alpha=gender, colour=gender, shape=gender)) +
    facet_grid(year ~ .) +
    xlab('Distance (km)') +
    ylab('Finish Difference (hrs)')
```

```
## `summarise()` has grouped output by 'distance'. You can override using the `.groups` argumen
## Warning: Using alpha for a discrete variable is not advised.
```

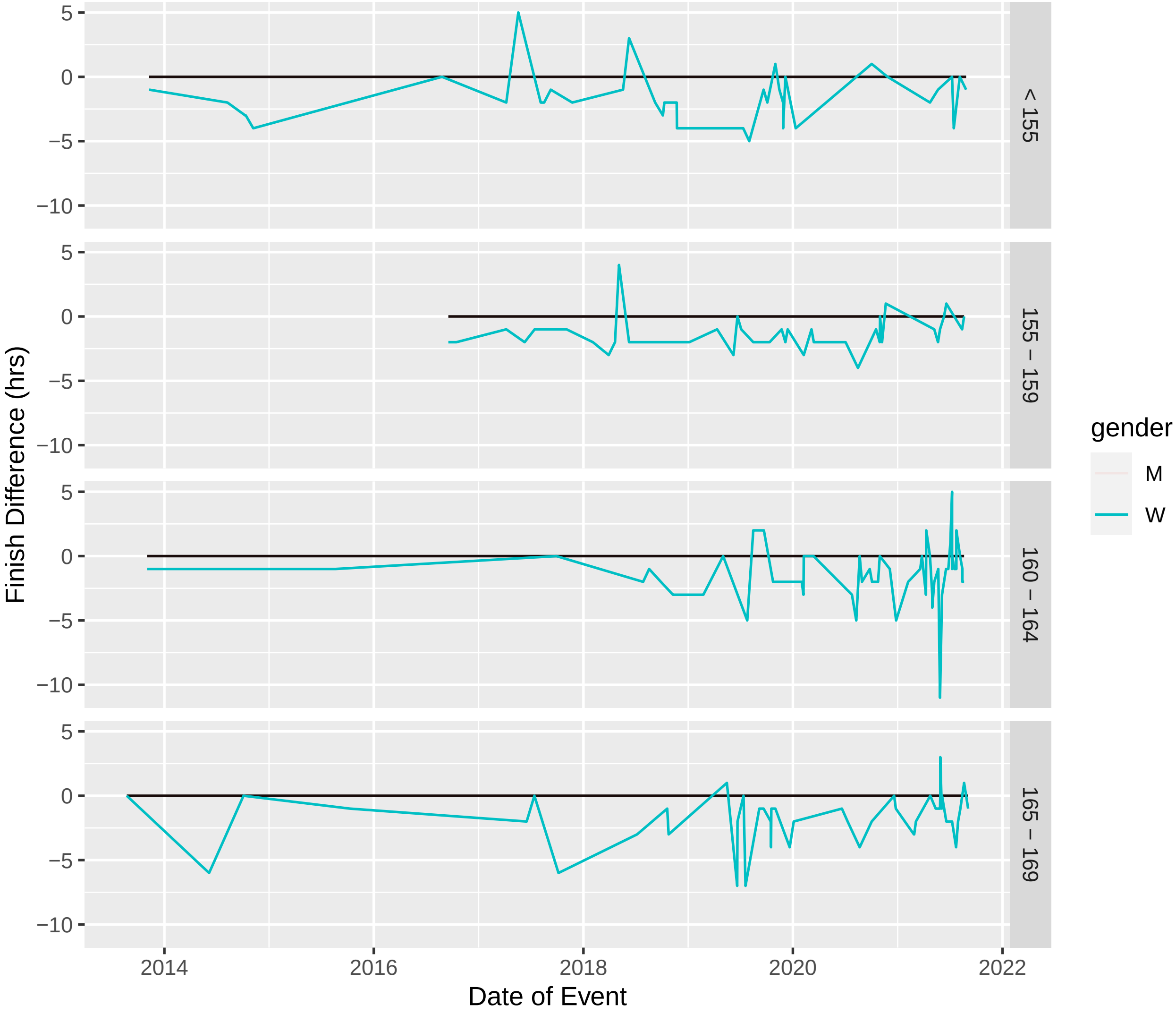


Women have shown continuous improvement

```
full_set %>%
  filter(gender %in% c('M', 'W') & distance >= 150) %>%
  group_by(distance, gender) %>%
  summarise(mean_time = mean(time_in_seconds, na.rm=TRUE),
            date = max(date)) %>%
  mutate(gender_diff = (mean_time[gender == 'M'] - mean_time) %/% 3600,
         distance_group = ifelse(distance >= 155 & distance < 160, '155 - 159',
                          ifelse(distance >= 160 & distance < 165, '160 - 164',
                          ifelse(distance >= 165 & distance < 170, '165 - 169',
                          ifelse(distance >= 175, '175+', '< 155'))))) %>%
  filter(!(distance_group %in% c('175+'))) %>%
  ggplot() +
    geom_line(aes(x=date, y=0)) +
    geom_line(aes(x=date, y=gender_diff, alpha=gender, colour=gender)) +
    facet_grid(distance_group ~ .) +
    xlab('Date of Event') +
    ylab('Finish Difference (hrs)')
```

```
## `summarise()` has grouped output by 'distance'. You can override using the `.groups` argumen
```

```
## Warning: Using alpha for a discrete variable is not advised.
```


< 155
Finish Difference (hrs)
Date of Event
gender
M
W