Homework #4

Andrea Navarrete (UNI: an2886)

NYC Marathon

In this project we want to analyse one of the most important events taking place every year in NYC, that is the NYC marathon where people all over the world come to run it.

Data

The data we are using comes from the New York Road Runners, where we are scrapping all the data for all runners, wheelcharis and handcycles that finished. At the TCS New York City Marathon 2018 that took place on november 4, 2018 we have:

Finishers by type of race:

- runners participants: 52,701
- wheelchair participants: 56
- handcycles participants: 52

The first approach for this analysis is to understand who are these particpants in terms of gender, age and where do they come from country / state.

(Note: 31 participants where drop out for incositencies during the parsing process during the scrapping)

```
library(tidyverse)
marathon <- read_csv('../final-project/marathon_2018.csv')
glimpse(marathon)
```

```
## Observations: 52,670
## Variables: 28
## $ bib                     <int> 1, 2, 3, 4, 5, 6, 9, 10, 11, 12, 14, 1...
## $ gender_age              <chr> "M25", "M22", "M28", "M27", "M26", "M2...
## $ gun_place               <int> 3, 2, 1, 4, 5, 8, 11, 18, 9, 13, 50, 6...
## $ gun_time                <time> 02:06:26, 02:06:01, 02:05:59, 02:08:3...
## $ name                    <chr> "Geoffrey Kamworor", "Shura Kitata", "...
## $ official_time           <time> 02:06:26, 02:06:01, 02:05:59, 02:08:3...
## $ pace_per_mile           <time> 04:50:00, 04:49:00, 04:49:00, 04:55:0...
## $ `percentile_age-graded` <chr> "97.25%", "97.57%", "97.6%", "95.69%",...
## $ place                   <chr> "Kapchorwa District", "Addis Ababa", "...
## $ `place_age-graded`      <dbl> 3, 2, 1, 4, 5, 9, 12, 6, 10, 14, 64, 7...
## $ `place_age-graded_of`   <dbl> 30581, 30581, 30581, 30581, 30581, 305...
## $ `place_age-group`       <dbl> 2, 1, 1, 3, 4, 2, 1, 1, 2, 3, 8, 1, 4,...
## $ `place_age-group_of`    <dbl> 2876, 773, 2876, 2876, 2876, 773, 4690...
## $ place_gender            <dbl> 3, 2, 1, 4, 5, 8, 11, 18, 9, 13, 46, 6...
## $ place_gender_of         <dbl> 30581, 30581, 30581, 30581, 30581, 305...
## $ place_overall           <dbl> 3, 2, 1, 4, 5, 8, 11, 18, 9, 13, 53, 6...
## $ place_overall_of        <dbl> 52697, 52697, 52697, 52697, 52697, 526...
## $ splint_10k              <time> 00:30:51, 00:30:48, 00:30:51, 00:30:4...
## $ splint_15k              <time> 00:45:49, 00:45:47, 00:45:48, 00:45:4...
## $ splint_20k              <time> 01:00:45, 01:00:39, 01:00:44, 01:00:4...
## $ splint_25k              <time> 01:15:44, 01:15:45, 01:15:45, 01:15:4...
## $ splint_30k              <time> 01:30:20, 01:30:20, 01:30:20, 01:30:2...
```

```
## $ splint_35k              <time> 01:45:19, 01:45:19, 01:45:19, 01:45:1...
## $ splint_40k              <time> 01:59:40, 01:59:50, 01:59:40, 02:01:2...
## $ splint_5k               <time> 00:15:43, 00:15:43, 00:15:43, 00:15:4...
## $ splint_half             <time> 01:03:59, 01:03:55, 01:03:57, 01:03:5...
## $ team                    <chr> "NIKE", "NIKE", "NIKE", "adidas", "adi...
## $ `time_age-graded`       <time> 02:06:26, 02:06:01, 02:05:59, 02:08:3...
```

Clean the Data

First filtering inconsistencies.

```
# Sanity Check
marathon_clean <- marathon %>%
  filter(is.na(splint_5k)  | is.na(splint_10k) | splint_5k <= splint_10k ) %>%
  filter(is.na(splint_10k) | is.na(splint_15k) | splint_10k <= splint_15k ) %>%
  filter(is.na(splint_15k) | is.na(splint_20k) | splint_15k <= splint_20k ) %>%
  filter(is.na(splint_20k) | is.na(splint_25k) | splint_20k <= splint_25k ) %>%
  filter(is.na(splint_25k) | is.na(splint_30k) | splint_25k <= splint_30k ) %>%
  filter(is.na(splint_30k) | is.na(splint_35k) | splint_30k <= splint_35k ) %>%
  filter(is.na(splint_35k) | is.na(splint_40k) | splint_35k <= splint_40k ) %>%
  filter(is.na(splint_half) | is.na(splint_20k) | splint_20k <= splint_half) %>%
  filter(is.na(splint_half) | is.na(splint_25k) | splint_half <= splint_25k)
```

Then, cleaning the columns and adding an identifier for the type of race they where participating.

Clean location

The location (city) in the data is autoreported, so there are different ways people could have enter this information the main patterns we saw in this columns are:

- National participantes reported: `City, state code`
- International participants reported their local city without the country

By spliting the city column into state, we have most of the complete information for national participants. For international participants we merge with a database of cities around the world provided by the library `maps`. This way we are able to fill country data for 76% of the participants which we hope to increase in the future.

```
library(maps)
data(world.cities)

repeated_cities <- world.cities %>%
  group_by(name) %>%
  summarise(n = n()) %>% filter(n>1) %>%
  select(name) %>% flatten_chr()

get_state_name <- function(abb){
  index <- match(abb, state.abb)
  return(state.name[index])
}

marathon_clean <- marathon_clean %>%
  separate(place, c('city', 'state'), sep=",") %>%
  separate(gender_age, c('gender','age' ), sep=1) %>%
  mutate(state = toupper(trimws(state)),
```

```
         state_name = get_state_name(state),
         city = gsub("[.]","", city),
         city = gsub("District","", city),
         city = trimws(city),
         age = as.numeric(age))

cities <- world.cities %>%
  filter(! name %in% repeated_cities) %>%
  rename(city = name) %>%
  select(city, country.etc)

marathon_clean <- marathon_clean %>%
  left_join(cities, by='city')

marathon_clean <- marathon_clean %>%
  mutate(country=ifelse(state %in% state.abb, 'USA', country.etc)) %>%
  select(-country.etc)

marathon_clean %>% group_by(country) %>% tally(sort=TRUE)
```

```
## # A tibble: 117 x 2
##    country         n
##    <chr>       <int>
##  1 USA         30967
##  2 <NA>        12185
##  3 Italy        1294
##  4 France        766
##  5 Netherlands   706
##  6 UK            629
##  7 Germany       618
##  8 China         479
##  9 Brazil        439
## 10 Canada        429
## # ... with 107 more rows
```

Once we have the `country` and the `state` for participants from the USA, it is possible to check and compare the number of participants by country for internationals (we filter by countries with at least 5 participants) or by state for national participants.

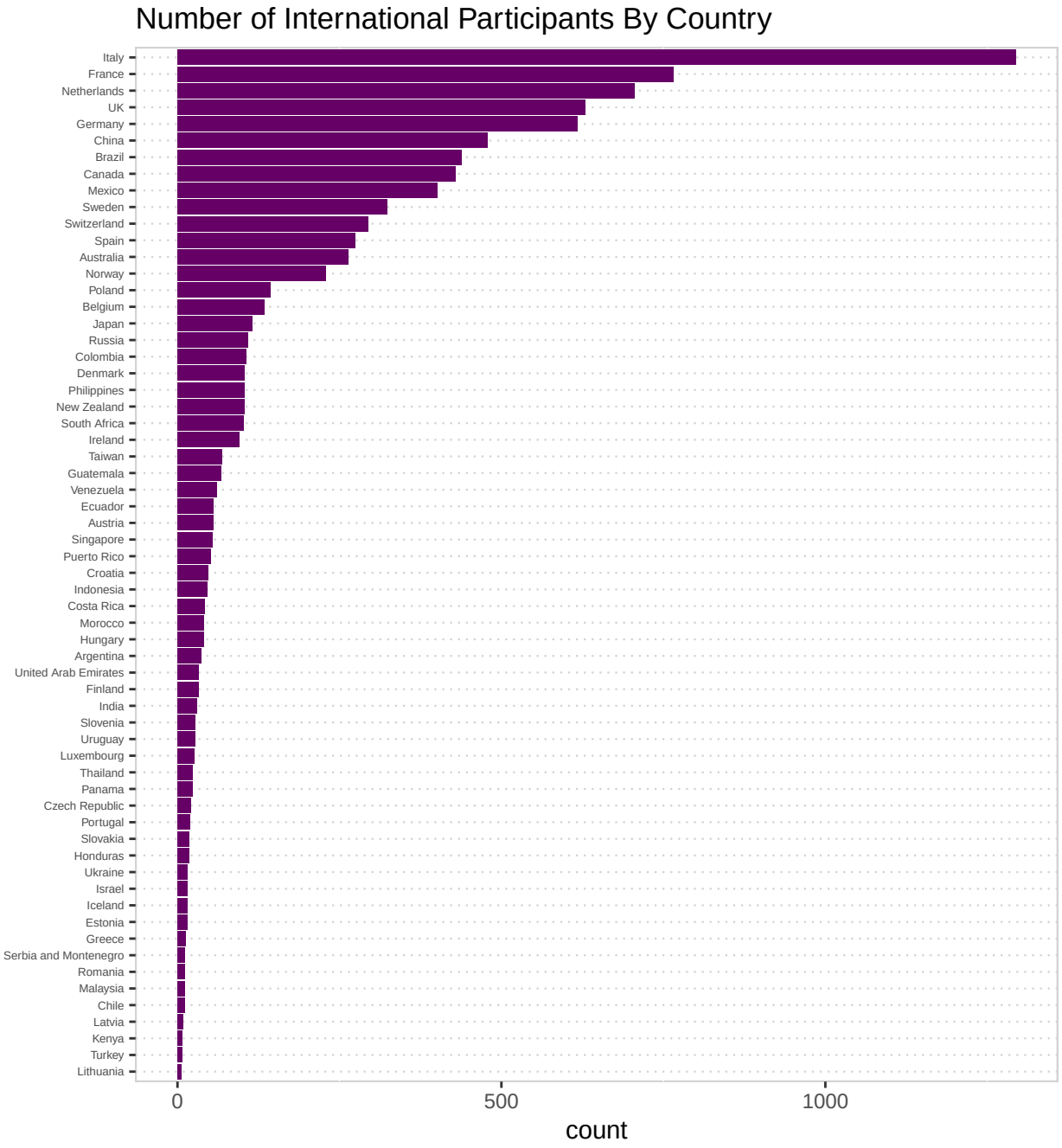
Number of international participants by country

```
# levels for ordering the countries for plot
levels_country <- marathon_clean %>%
          filter(type == "R" & !is.na(country)) %>%
          group_by(country) %>% summarise(n=n()) %>%
          arrange(n) %>%
  select(country) %>% flatten_chr()

# Filter by countries with at least 5 participants
countrys_min_1 <- marathon_clean %>%
          filter(type == "R" & !is.na(country)) %>%
          group_by(country) %>%
          summarise(n = n()) %>%
```

```
          filter(n > 5) %>% select(country) %>% flatten_chr()

# Plot
marathon_clean %>%
  filter(type == "R" & !is.na(country) & country != 'USA' & country %in% countrys_min_1) %>%
  mutate(country =  factor(country, levels = levels_country)) %>%
  ggplot(aes(x=country)) +  geom_bar(fill = '#660066') +
  coord_flip() +
  ggtitle("Number of International Participants By Country") +
  theme(panel.background = element_rect(fill = "white", colour = "lightgray"),
  panel.grid.major.x = element_blank() ,
  panel.grid.major.y = element_line(linetype=3, color="lightgray", size=0.4), axis.text.y = element_tex
  axis.title.y = element_blank())
```

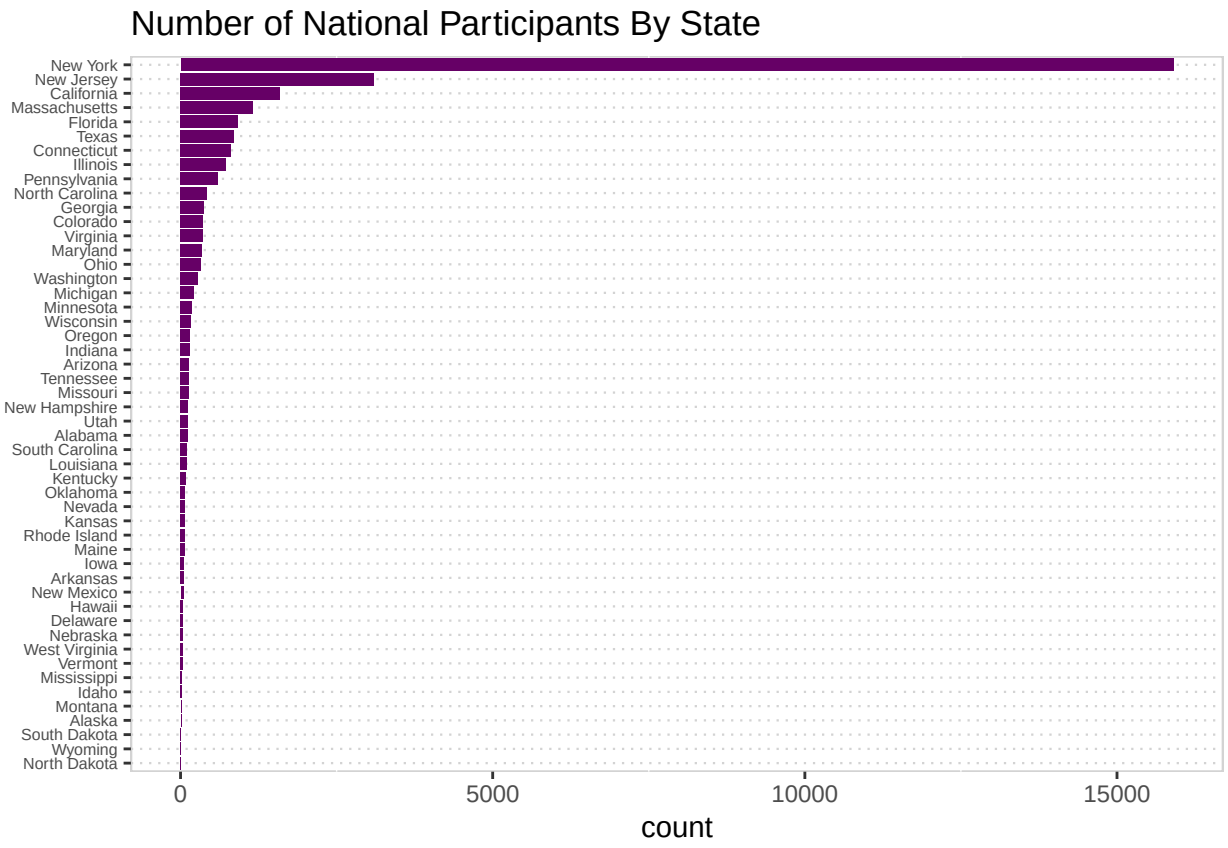


Number of national participants by state

```
# levels for ordering the state for plot
levels_state <- marathon_clean %>%
          filter(type == "R" & !is.na(state_name)) %>%
          group_by(state_name) %>% summarise(n=n()) %>%
          arrange(n) %>%
  select(state_name) %>% flatten_chr()

# Plot
```

```
marathon_clean %>%
  filter(type == "R" & !is.na(state_name)) %>%
  mutate(state_name =  factor(state_name, levels = levels_state)) %>%
  ggplot(aes(x=state_name)) +  geom_bar(fill = '#660066') +
  coord_flip() +
  ggtitle("Number of National Participants By State") +
  theme(panel.background = element_rect(fill = "white", colour = "lightgray"),
  panel.grid.major.x = element_blank() ,
  panel.grid.major.y = element_line(linetype=3, color="lightgray", size=0.4), axis.text.y = element_tex
  axis.title.y = element_blank())
```



From this graphs below we can notice where the participants come from. Apart from the US, most of them come from Europe (Italy, France, Uk, Germany), North America (Canada, Mexico) but there is also great participation from China and Brasil, but no much participation from countries in Africa which are the ones that have usually won this marathon.

On the other hand, by analysing for states, as we expected we can see that most of the participants come from NY and NJ which is where the marathon takes place. It is importance to notice that apart from this two states, the participation can't be associated by distance to the states being California, Massachussets, Florida and Texas the following states with greater participation.

Ratio between Female / Male

Another measure we are interested is to see if the representation of woman/men across each country is similar. To do so, we careated a new variable which measures the ratio between gender. This way countries/states with ratio close to one, have similar representation.

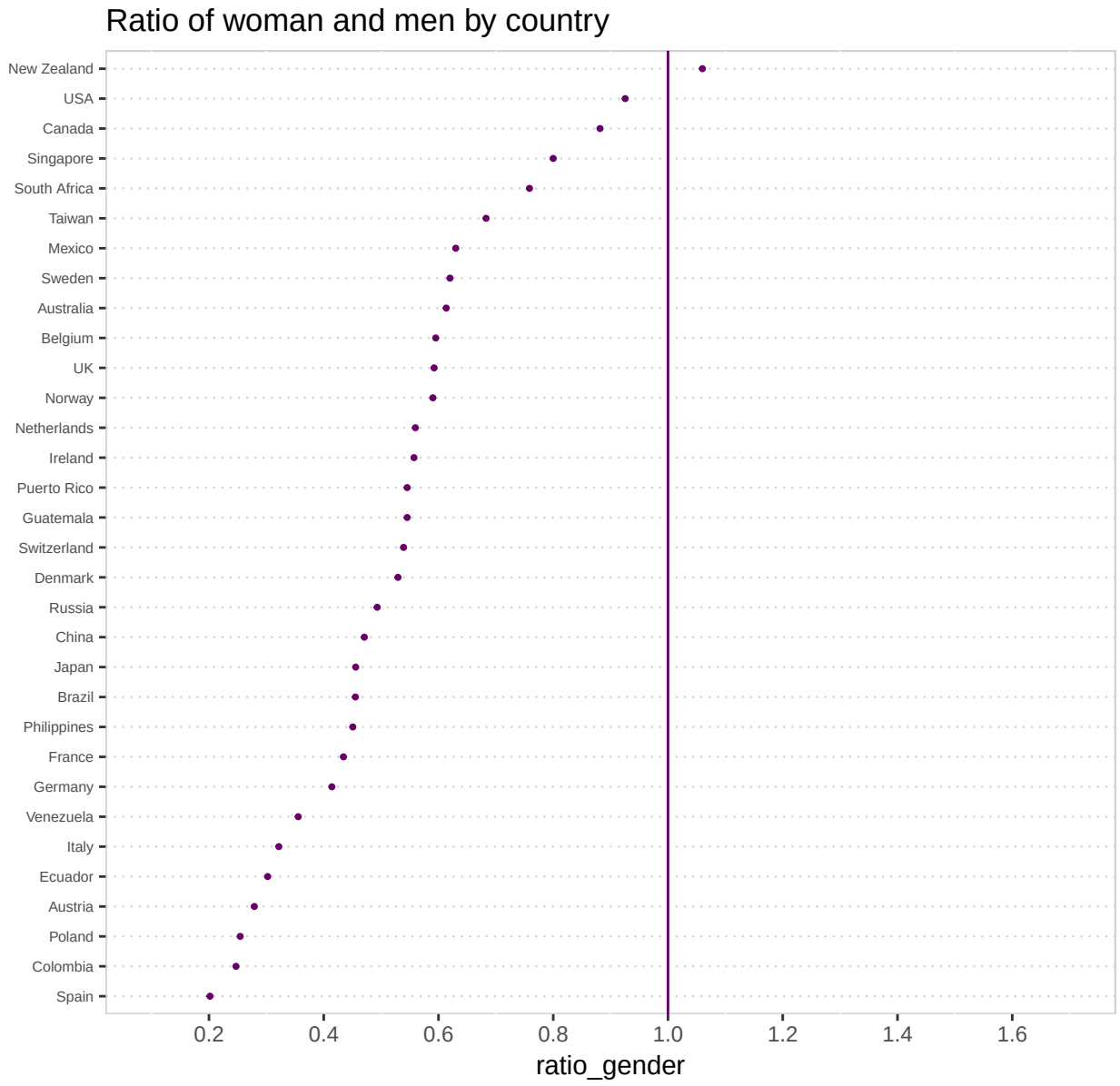
The total ratio across all runners is around 0.72 :

```
runners <- marathon_clean %>%
  filter(type=='R' & !is.na(gender))
sum(runners$gender == 'F') / sum(runners$gender == 'M')
```

```
## [1] 0.7242692
```

If we analyse by country for international participants:

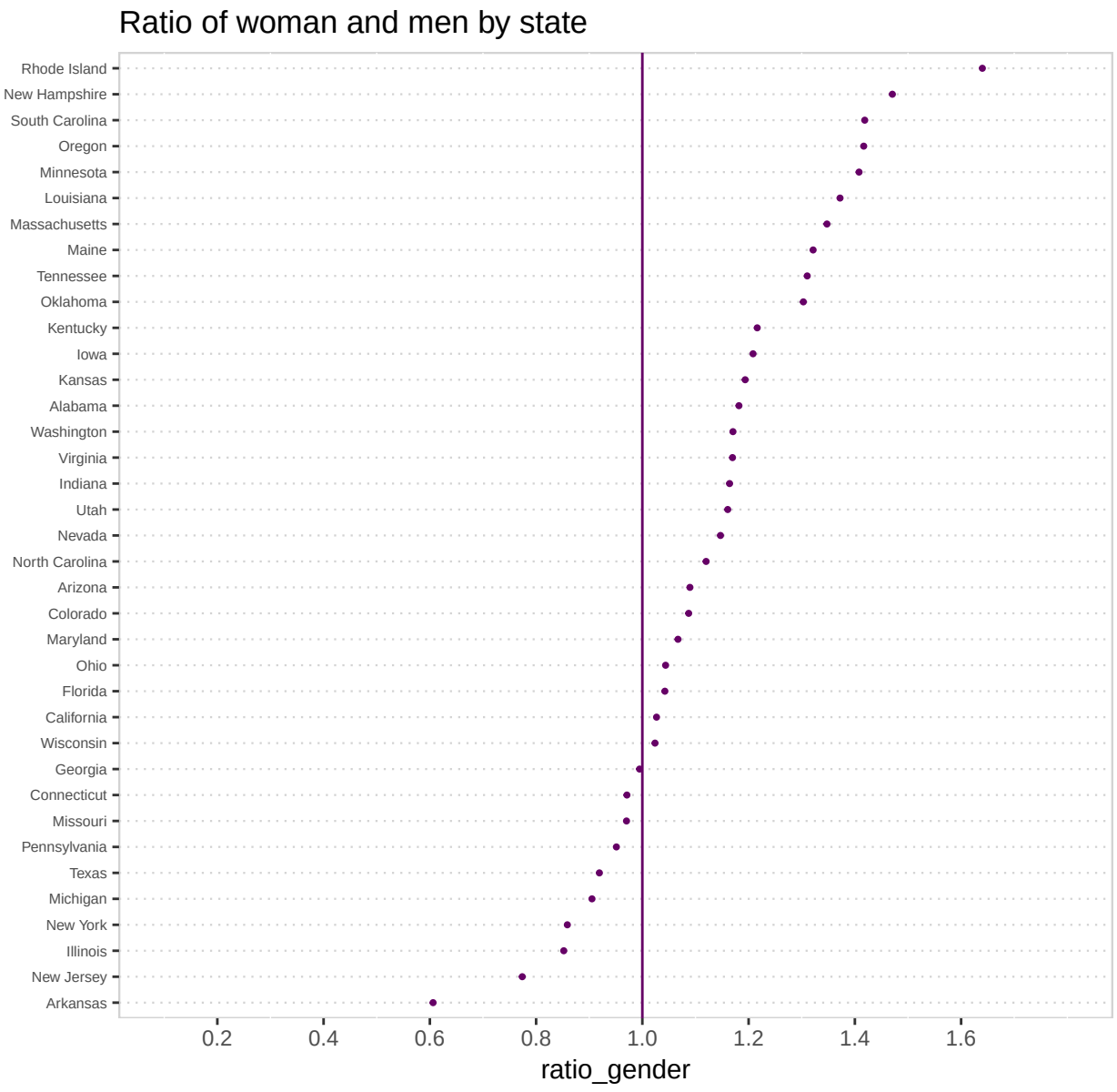
```
# For runners
marathon_clean %>%
  filter(type == 'R') %>%
  filter(country != 'NA') %>%
  group_by(country) %>%
  summarise(n = n(),
            num_male = sum(gender == 'M'),
            num_female = sum(gender == 'F'),
            ratio_gender = num_female / num_male) %>%
  filter(n > 50) %>%
  arrange(desc(ratio_gender)) %>%
  ggplot(aes(x = ratio_gender, y = fct_reorder(country, ratio_gender))) +
  geom_point(color='#660066', size=0.7) +
  geom_vline(xintercept=1, color='#660066') +
  ggtitle("Ratio of woman and men by country") +
  scale_x_continuous(breaks = round(seq(0, 1.7, by = 0.2),1), limits=c(0.1, 1.7)) +
  theme(panel.background = element_rect(fill = "white", colour = "lightgray"),
  panel.grid.major.x = element_blank() ,
  panel.grid.major.y = element_line(linetype=3, color="lightgray", size=0.4), axis.text.y = element_tex
  axis.title.y = element_blank())
```



And by state for national participants:

```
marathon_clean %>%
  filter(type == 'R') %>%
  filter(state_name != 'NA') %>%
  group_by(state_name) %>%
  summarise(n = n(),
            num_male = sum(gender == 'M'),
            num_female = sum(gender == 'F'),
            ratio_gender = num_female / num_male) %>%
  filter(n > 50) %>%
  arrange(desc(ratio_gender)) %>%
  ggplot(aes(x = ratio_gender, y = fct_reorder(state_name, ratio_gender))) +
  geom_point(color='#660066', size=0.7) +
  geom_vline(xintercept=1, color='#660066') +
  ggtitle("Ratio of woman and men by state") +
```

```
  scale_x_continuous(breaks = round(seq(0, 1.7, by = 0.2),1), limits=c(0.1, 1.8)) +
  theme(panel.background = element_rect(fill = "white", colour = "lightgray"),
  panel.grid.major.x = element_blank() ,
  panel.grid.major.y = element_line(linetype=3, color="lightgray", size=0.4), axis.text.y = element_tex
  axis.title.y = element_blank())
```



It is possible to notice that for almost all of the countries there is more participation of men vs woman, nevertheless it is interested to notice that for national participants this is not the case many states have more woman than men participating. From this graphs we can see the importance of separating by country/state the ratio, since NY has most of the participants the overall ratio behaves very similar to NY ratio which only tell us about the local participants but not from the rest of the participants coming from across the US and the rest of the countries.

Perfomance by location

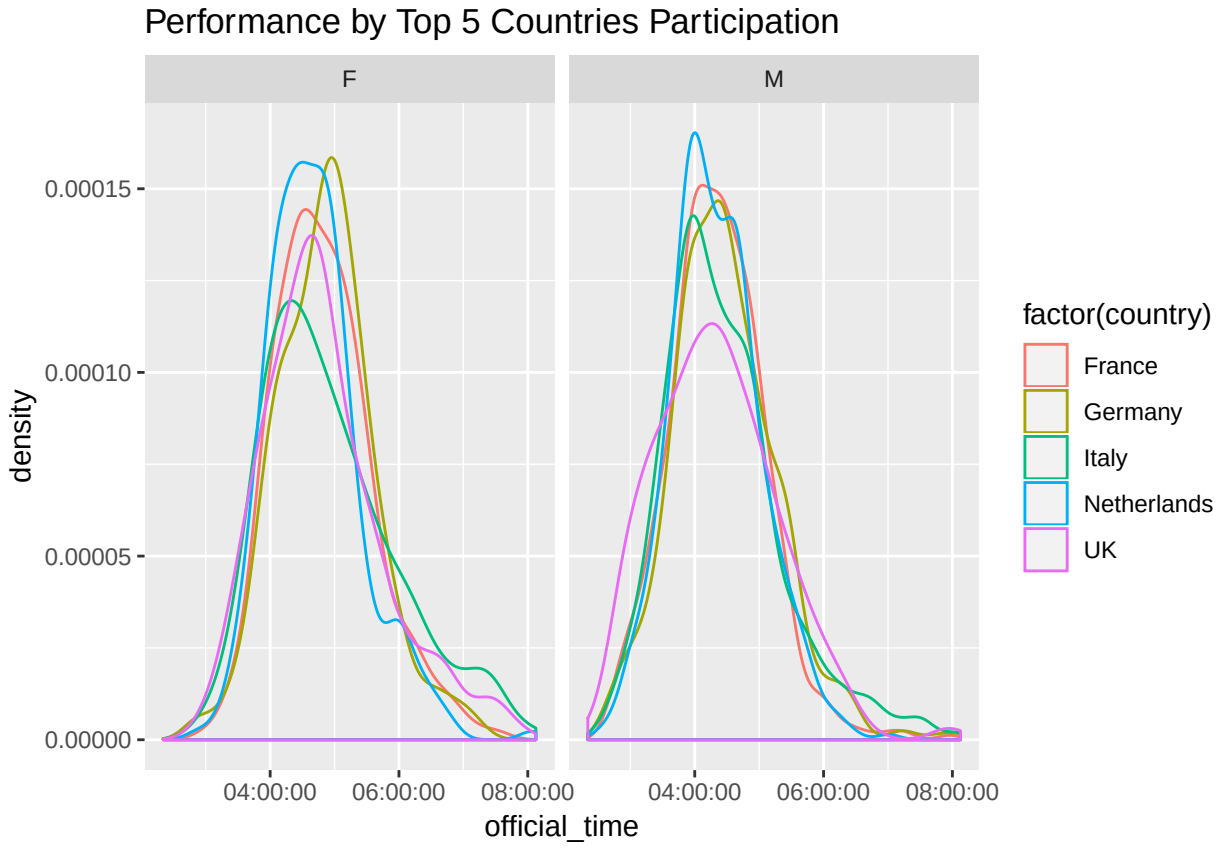
```
top_5_countries <- marathon_clean %>%
  filter(type == 'R') %>%
  filter(country != 'NA' & country != 'USA') %>%
  group_by(country) %>%
  tally(sort=TRUE) %>% top_n(5) %>%
  select(country) %>% flatten_chr()
```

```
## Selecting by n
```

```
top_5_countries
```

```
## [1] "Italy"       "France"      "Netherlands" "UK"          "Germany"
```

```
marathon_clean %>%
  filter(type == 'R') %>%
  filter(country %in% top_5_countries) %>%
  ggplot(aes(x=official_time, color = factor(country))) +
  geom_density() + facet_grid(~gender) +
  ggtitle("Performance by Top 5 Countries Participation")
```



In this graph we are able to compare the distribution of the total time for the top 5 countries in terms of representation of participants, dividing the graph for women and men. Although the distributions behave similar, it is possible to see some small differences specially for men in Italy, having a smaller mode, while Germany having greater one. For women what is interested is the variance for UK seems a little greater, while Italy and Netherlands have smaller one.

Next steps

Some of the next steps would be:

- Compared participation accross multiple years of Marathons (for location and gender ratio)
- Analysing ages by location
- Analysing performance by location