U.S. High School Track & Field and Cross-Country Data Analysis

Data Science 2019 Fall Final Project

Wei-Hua Hsu (Wafer)

2019-12-12

Introduction

In the United States, sports is an important part of American culture. American football is the most popular sport to watch, but running, jogging, and trail running is the most popular exercise people practice daily. Between the U.S. high schools, there are thousands of Track & Field and Cross Country competitions held within colleges and high schools every year. For those students who are dedicated to sports, the performance list is crucial since they would go to college depending on their performance records. However, in the view of colleges, they prefer to recruit new students who are from a middle class or lower income family. In this way the college can save budgets by paying their athletes through the sponsorship from the U.S. government. Under this premise, this study aims to find out which states tend to have more elite athletes. In addition to personal training and genetic strength, many studies indicate the environmental factors can have a significant effect on athlete's performance. Therefore, I plan to explore the relationship between player performance and climate factors (rainfall, temperature, sun hours, etc.) among states, and then apply the possible results to statewide family household income.

Data Collecting

This project will include four main aspects of data:
1. The **information of the U.S.**: 50 states and District of Columbia, with state's latitude and longitude.
2. The performance list of **top 500 high school athletes**: the top 500 U.S. high school athletes in the sports of Track & Field and Cross Country (XC) in 2018. The datasets are separated by gender, there are three different events 800m, 1600m, 3200m for indoor/outdoor Track & Field (ITF and OTF), and the XC comes with 5k performance list. The athlete's hometown and the time he/she finished the race are included.

Indoor Track & Field	Outdoor Track & Field	Cross Country
boys_800m	boys_800m	boys_XC_5k
girls_800m	boys_800m	girls_XC_5k
boys_1600m	boys_1600m	
girls_1600m	girls_1600m	
boys_1600m	boys_3200m	
girls_1600m	girls_3200m	

3. **U.S. Climate data**: all statewide climate data are classified by year, including temperature, rainfall inches, humidity, sunshine hours, wind speed, and elevations.

Climate Data
average *temperature*
average *precipitation*
morning and afternoon *humidity*
annual *sunshine hours*
average *wind speed*
highest/lowest *elevations*

4. **U.S. Social Economy data**: median household income by state (2013-2017).

Study Questions

1. What does the statewide player performance looks like? Are there some states tend to have better performance?

2. Does climate matter? How does the climate factors (temperature, rainfall inches, humidity, sunshine hours, wind speed, elevations) relate to elite athletes in the U.S.?
3. How wealthy are those outstanding athletes? Is there any lower-income state that tends to have a better performance of athletes?

Overview of results

1. The indoor/outdoor Track & Field list shows different relationship between athletes with different states.

2. People from northeastern, west coast (WA, CA), and southern (TX, FL) of the United States engage more in sports, and the western inland states of Colorado and Utah are second popular with sports.

3. Climate and natural conditions do affect players' performance in a certain aspect.

4. Female athletes are more spread out between states than male athletes.

5. Natural phenomenons such as the population from each state need to be taken into consideration.

6. The Cross Country results are evenly distributed.

Data import, cleaning, and tidying

1. State information in the U.S.

- I found a data file with each state's information. But I do not need "medals" info and this data miss West Virginia and South Dakota, I tidy and add more info on it.

```
# import
state_info <- read_csv("./data/state_info.csv")
```

```
## Parsed with column specification:
## cols(
##   state = col_character(),
##   medals = col_double(),
##   location = col_character(),
##   lat = col_double(),
##   lon = col_double()
## )
```

```
# tidy
state_info <- state_info %>%
  select(-medals) %>%
  bind_rows(list(state = c("WV", "SD"),
                 location = c("West Virginia", "South Dakota"),
                 lat = c(39.000000, 44.500000),
                 lon = c(-80.500000, -100.000000)))
# dataset - state_info
```

- To draw a U.S. map, I found an useful package "usmap", this dataset has a more detailed position information, and it helps me with graphing the map.

```r
library(usmap)
# tidy & save
usmap <- map_data("state") %>%
  select(1, 2, 3, 5)
# dataset - usmap
```

2. The performance list of top 500 high school athletes

- **Function**: Since the format of high school data are very similar and those are embedded in an excel file. I write 2 functions to tidy the data, one is for tiding the data, the other is to convert the time format to seconds. Then applying `map()` function on the data.

```r
## function1 - tidy High School data
tidy_hs <- function(x) {
  x %>%
    slice(which(row_number() %% 2 == 0)) %>%
    mutate(state = str_extract(`ATHLETE/TEAM`, "(^(?i)[a-z][a-z])")) %>%
    separate(`ATHLETE/TEAM`, into = c("Empty", "Team"), sep = "(^(?i)[a-z][a-z])") %>%
    select(8, 5) %>%
    mutate(state = str_to_upper(.$state)) %>%

    # column bind all the second row of data
    cbind(
      x %>%
        slice(which(row_number() %% 2 == 1)) %>%
        select(-3) %>%
        set_names(c("Rank", "Time", "Athlete", "Grade", "Meet/Place")) %>%
        mutate(Place = str_extract(`Meet/Place`, "(\\d[a-z][a-z])$")) %>%
        mutate(`Meet/Place` = str_replace(`Meet/Place`, "(\\d[a-z][a-z])$", ""))
      ) %>%
    set_names(c("state", "Team", "Rank", "Time",
                "Athlete", "Grade", "Meet", "Place")) %>%
    select(3, 5, 4, 1, 2, 6, 7, 8) %>%

    # join the state info
    left_join(state_info, by = "state")
}
```

```r
## function2 - convert time to second
count_sec <- function(x) {
  x <- x %>%
    separate(Time, into = c("minute", "second"), sep = ":")

  # set the time variable as numeric
  x$minute <- as.numeric(x$minute)
  x$second <- as.numeric(x$second)

  # calculate the time variable
  x <- x %>%
    mutate(Time = (minute * 60) + second) %>%
```

```
    select(1, 2, 13, 5, 10:12, 6:9)

  return(x)
}
```

- **Data import**: There are three data files for top 500 high school athletes: 1) indoor Track & Field (*hs_indoor18*); 2) outdoor Track & Field (*hs_outdoor18*); 3) Cross Country (*hs_xc18*). These are high school ranking results with national meeting events in the year of 2018.
  Since the codes are basically the same, the code is shown on .pdf file only includes ***hs_indoor18***

```
hs_indoor18 <- "./data/HS_indoor18.xlsx"
# read excel
hs_indoor18 <- hs_indoor18 %>%
  excel_sheets() %>%
  purrr::set_names() %>%
  map(read_excel, path = hs_indoor18)
# tidy & convert the time format
hs_indoor18 <- map(hs_indoor18, tidy_hs)
hs_indoor18 <- map(hs_indoor18, count_sec)

# not sure why but this variable couldn't change to numeric automatically
hs_indoor18$boys_1600m$Rank <- as.numeric(hs_indoor18$boys_1600m$Rank)
```

- Take a glance at the high school performance list, the format of the rest of datasets are basically the same.

```
head(hs_indoor18$boys_800m, 10)
```

```
##    Rank         Athlete   Time state       location      lat        lon
## 1     1       JOSH HOEY 107.67    PA   Pennsylvania 40.86020  -77.83862
## 2     2    JETT CHARVET 110.91    CA     California 36.53154 -119.58617
## 3     3     IAN DELGADO 111.06    NC North Carolina 35.38736  -78.45506
## 4     4    LUIS PERALTA 111.74    NJ     New Jersey 39.66502  -74.73821
## 5     5      SEAN DOLAN 111.77    NJ     New Jersey 39.66502  -74.73821
## 6     6     JASON GOMEZ 111.86    CA     California 36.53154 -119.58617
## 7     7    JAKE MERRELL 112.33    TX          Texas 31.03097  -98.32633
## 8     8 ALFRED CHAWONZA 112.43    NJ     New Jersey 39.66502  -74.73821
## 9     9  SCOTT THOMPSON 112.48    TN      Tennessee 35.82024  -86.34376
## 10   10   MATTHEW RIZZO 112.55    NY       New York 42.16573  -74.94805
##                      Team Grade
## 1         Bishop Shanahan  2018
## 2      Heritage High (NC)  2018
## 3              Green Hope  2018
## 4              Passaic HS  2019
## 5      Hopewell Valley HS  2019
## 6      Westmont High (CC)  2018
## 7           Turkey Valley  2018
## 8     St. Benedict's Prep  2019
## 9   Brentwood High School  2018
## 10            Bronxville  2019
##                                                 Meet Place
```

```
## 1                            Boston University Last Chance   2nd
## 2                             New Balance Nationals Indoor   1st
## 3                             New Balance Nationals Indoor   2nd
## 4                             New Balance Nationals Indoor   3rd
## 5                             New Balance Nationals Indoor   4th
## 6                             New Balance Nationals Indoor   5th
## 7                             New Balance Nationals Indoor   6th
## 8                                    Fastrack Last Chance   1st
## 9  Tennessee State HS Indoor Track and Field Championships   1st
## 10                            New Balance Nationals Indoor   7th
```

```
head(hs_outdoor18$girls_800m, 10)
```

```
##    Rank             Athlete   Time state      location      lat        lon
## 1     1     CAITLIN COLLIER 120.85    FL       Florida 27.97762  -81.76961
## 2     2           ATHING MU 124.51    NJ    New Jersey 39.66502  -74.73821
## 3     3   CATHILYN MCINTOSH 125.22    CA    California 36.53154 -119.58617
## 4     4 GABRIELLE WILKINSON 125.72    PA  Pennsylvania 40.86020  -77.83862
## 5     5   VICTORIA VANRIELE 125.99    NJ    New Jersey 39.66502  -74.73821
## 6     6    SAMANTHA FRIBORG 126.13    MA Massachusetts 42.35875  -71.53148
## 7     7       MICHAELA ROSE 126.35    VA      Virginia 37.73086  -78.37665
## 8     8       BROOKE MANSON 127.16    WA    Washington 47.37227 -120.59234
## 9     9   VICTORIA STARCHER 127.42    WV West Virginia 39.00000  -80.50000
## 10   10         GRACE BOONE 127.46    VA      Virginia 37.73086  -78.37665
##                                                  Team Grade
## 1                                          Bolles HS  2018
## 2                                         Trenton TC  2020
## 3                                    Del Oro High (SJ)  2018
## 4                                     Friends' Central  2018
## 5                                Governor Livingston HS  2020
## 6                           Acton-Boxborough High School  2018
## 7   Faith In Action Ministries Athletics and Recreation  2021
## 8                                  Eastlake High School  2018
## 9                                             Ripley  2020
## 10                                      Pulaski County  2019
##                                        Meet Place
## 1  16th Annual Music City Distance Carnival   4th
## 2              New Balance Nationals Outdoor   1st
## 3   CIF State Track and Field Championships   1st
## 4              New Balance Nationals Outdoor   2nd
## 5              New Balance Nationals Outdoor   3rd
## 6              New Balance Nationals Outdoor   4th
## 7              New Balance Nationals Outdoor   5th
## 8                     Brooks PR Invitational   2nd
## 9              New Balance Nationals Outdoor   6th
## 10             New Balance Nationals Outdoor   7th
```

3. The U.S. Climate data

- For climate data, I do a lot of web scraping as follow.

- Average annual temperature by state

```
# webscraping
temp_url <- read_html("https://www.currentresults.com/Weather/US/average-annual-state-temperat
temperature <- html_table(temp_url, fill = T)

# tidy & join state_info
temperature <- rbind(temperature[[1]], temperature[[2]], temperature[[3]]) %>%
  set_names(c("location", "avg_F", "avg_C", "Rank")) %>%
  left_join(state_info, by = "location")

# dataset - temperature
```

- Average annual precipitation by state

```
# webscraping
rain_url <- read_html("https://www.currentresults.com/Weather/US/average-annual-state-precipit
rainfall <- html_table(rain_url, fill = T)

# tidy & join state_info
rainfall <- rbind(rainfall[[1]], rainfall[[2]], rainfall[[3]]) %>%
  set_names(c("location", "Inches", "Millimeters", "Rank")) %>%
  left_join(state_info, by = "location")

# dataset - rainfall
```

- Average annual morning and afternoon humidity (%) by states: Since the row with "Connecticut" and "Massachusetts" contain an unreadable UTF-8 signs, I duplicate the info and delete the former one.

```
# webscraping
humid_url <-
  read_html("https://www.currentresults.com/Weather/US/annual-average-humidity-by-state.php")
humidity <- html_table(humid_url, fill = T)

# tidy & join state_info
humidity <- rbind(humidity[[1]], humidity[[2]], humidity[[3]]) %>%
  set_names(c("location", "place", "morning", "afternoon")) %>%
  # fix the imput in the row with "location = Connecticut"
  filter(place != "Hartford" & place != "Boston") %>%
  bind_rows(list(location  = c("Connecticut", "Massachusetts"),
                 place     = c("Hartford", "Boston"),
                 morning   = c(79, 75),
                 afternoon = c(52, 59))
            ) %>%
  # join state_info
  left_join(state_info, by = "location")

# dataset - humidity
```

- Average annual sunshine hours by states

```r
# webscraping
sun_url <-
  read_html("https://www.currentresults.com/Weather/US/average-annual-state-sunshine.php")
sunshine <- html_table(sun_url, fill = T)

# tidy & join state_info
sunshine <- rbind(sunshine[[1]], sunshine[[2]], sunshine[[3]]) %>%
  mutate(location = State) %>%
  select(6, everything(), -1) %>%
  mutate(`% Sun` = as.integer(`% Sun`)) %>%
  mutate(`Total Hours` = as.integer(`Total Hours`)) %>%
  left_join(state_info, by = "location")

# dataset - sunshine
```

- Average Wind Speed by states (with the U.S. Population data): Because the `comma` and the `dash` are read as a "character", it needs more code to change the type of continuous variable.

```r
# webscraping
windsp_url <-
  read_html("http://www.usa.com/rank/us--average-wind-speed--state-rank.htm")
windspeed <- html_table(windsp_url, fill = T)[[2]]

# tidy & join state_info
windspeed <- windspeed %>%
  set_names(c("Rank", "avg_WindSpeed", "location / Population")) %>%
  slice(2:nrow(windspeed)) %>%
  separate("location / Population", into = c("location", "Population"), sep = " / ") %>%
  mutate(Rank = str_replace_all(.$Rank, "\\D", "")) %>%
  mutate(avg_WindSpeed = str_extract(.$avg_WindSpeed, "\\d\\d.\\d\\d")) %>%
  mutate(Population = str_replace_all(.$Population, "\\D", "")) %>%
  mutate(avg_WindSpeed = as.numeric(avg_WindSpeed)) %>%
  mutate(Population = as.numeric(Population)) %>%
  mutate(Rank = as.numeric(Rank)) %>%
  left_join(state_info, by = c("location"))

# dataset - windspeed
```

- Elevations by states: Because the `comma` and the `dash` are read as a "character", it needs more code to change the type of continuous variable.

```r
# webscraping
elev_url <-
  read_html("https://www.infoplease.com/world/united-states-geography/highest-lowest-and-mean-
elevation <- html_table(elev_url, fill = T)[[1]]

# tidy & join state_info
elevation <- elevation %>%
  set_names(c("location",      "avg_Elevation",
              "Highest Point", "Highest Elevation",
              "Lowest Point",  "Lowest Elevation")) %>%
```

```r
  mutate(avg_Elevation = str_replace_all(.$avg_Elevation, ",", "")) %>%
  mutate(avg_Elevation = as.numeric(avg_Elevation)) %>%
  mutate(`Highest Elevation` = str_replace_all(.$`Highest Elevation`, ",", "")) %>%
  mutate(`Highest Elevation` = as.numeric(`Highest Elevation`)) %>%
  mutate(`Lowest Elevation` = str_replace_all(.$`Lowest Elevation`, "Sea level", "0")) %>%
  mutate(`Lowest Elevation` = str_replace_all(.$`Lowest Elevation`, ",", "")) %>%
  mutate(`Lowest Elevation` = str_replace_all(.$`Lowest Elevation`, "\\D", "-")) %>%
  mutate(`Lowest Elevation` = as.numeric(`Lowest Elevation`)) %>%
  mutate(location = str_replace(.$location, "D.C.", "District of Columbia")) %>%
  left_join(state_info, by = c("location")) %>%
  na.omit()

# dataset - elevation
```

4. The U.S. Social Economy data

- Median Household Income by State (2013-2017)

```
# read .csv file
median_income <- read_csv("./data/Median_Income.csv")

# tidy
median_income <- median_income %>%
  set_names(c("location", "Income", "Margin of Error")) %>%
  slice(3:nrow(median_income)-1) %>%
  mutate(Income = str_replace_all(.$Income, "\\D", "")) %>%
  mutate(Income = as.numeric(Income)) %>%
  mutate(`Margin of Error` = str_replace_all(.$`Margin of Error`, "\\D", "")) %>%
  mutate(`Margin of Error` = as.numeric(`Margin of Error`)) %>%
  left_join(state_info, by = "location")

# dataset - median_income
```

Question 1

What does the statewide player performance looks like? Are there some states tend to have better performance?

- Write a function to get the subset of performance data

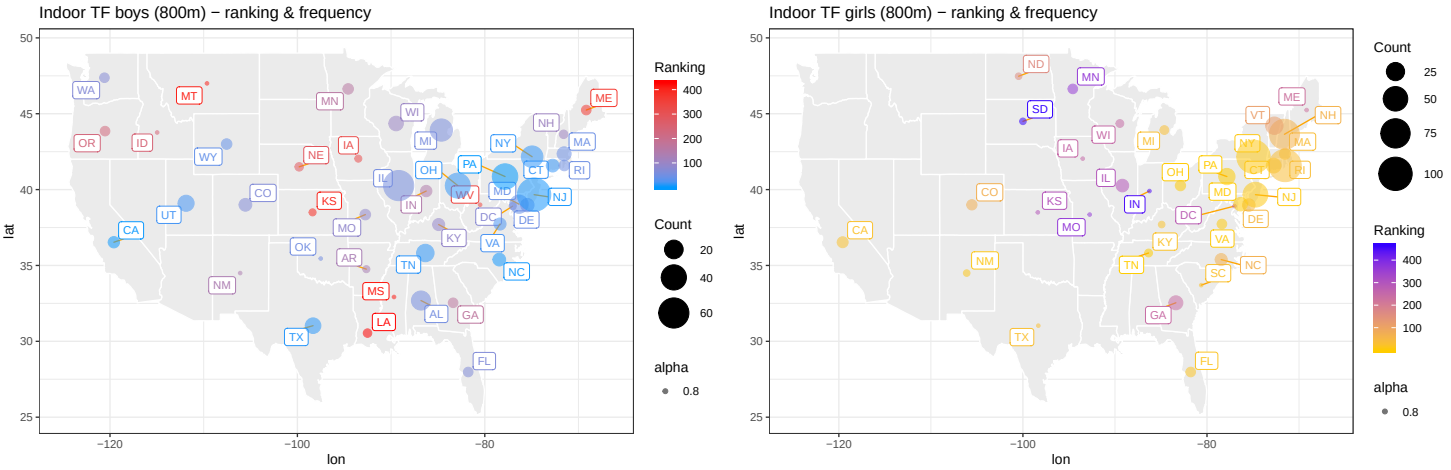
```
## function3 - subset data with the frequency of state
sub_data <- function(x) {
  x %>%
    # group by state
    group_by(state) %>%
    # aggregate the amount of athletes
    count() %>%
    left_join(x, by = "state") %>%
    group_by(state) %>%
    # keep the highest ranking of athletes by each state
    filter(Rank == min(Rank))
}

sub_otf <- map(hs_outdoor18, sub_data)
sub_itf <- map(hs_indoor18, sub_data)
sub_xc <- map(hs_xc18, sub_data)

# the format of sub_otf / sub_itf / sub_xc are basically the same
# take a glance at one of each
```

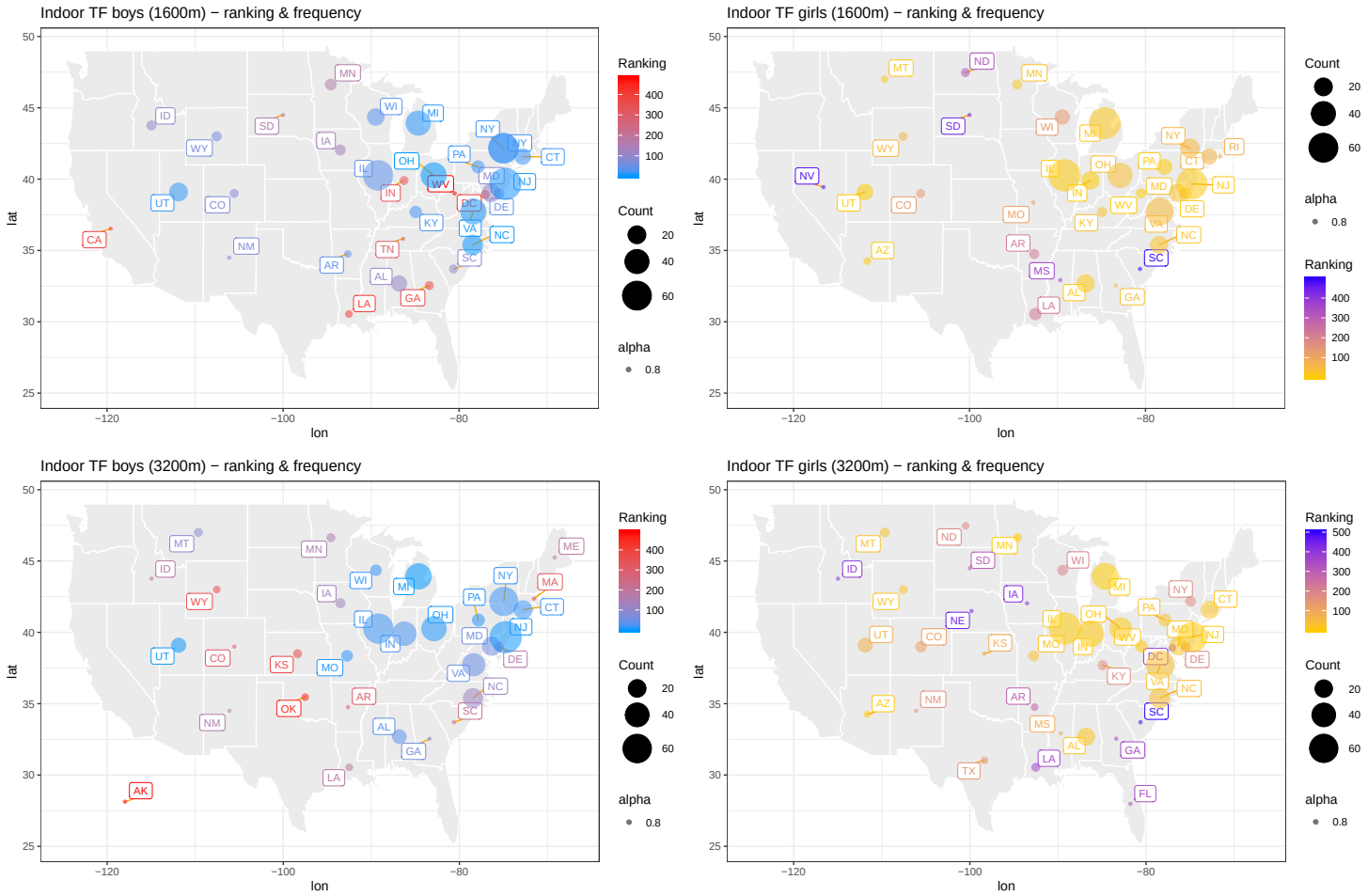
- Visualize the amount of athletes from each state, the bigger the "count circle" is, the greater number of athletes' hometowns are. And the color gradient records the best ranking a state's athlete got. The results are separate by different event and gender.

- Indoor Track & Field

```r
# size = frequency by states
# rank = the best ranking athlete from the state

# ----- Indoor TF 800m ----- #

ggplot(sub_itf$boys_800m, aes(x = lon, y = lat, color = Rank)) +
  geom_polygon(data = usmap, aes(x = long, y = lat, group = group),
               color = "white", fill = "grey92") +
  # label states
  ggrepel::geom_label_repel(aes(label = state), data = sub_itf$boys_800m,
    size = 3, label.size = 0, segment.color = "orange") +
  # point size & ranking
  geom_point(aes(size = n, color = Rank, alpha = 0.8)) +
  scale_color_continuous("Ranking", low = "#0099FF", high = "red") +
  scale_size_continuous("Count", range = c(1, 12)) +
  labs(title = "Indoor TF boys (800m) - ranking & frequency") -> itf_boys_800m

ggplot(sub_itf$girls_800m, aes(x = lon, y = lat, color = Rank)) +
  geom_polygon(data = usmap, aes(x = long, y = lat, group = group),
               color = "white", fill = "grey92") +
  # label states
  ggrepel::geom_label_repel(aes(label = state), data = sub_itf$girls_800m,
    size = 3, label.size = 0, segment.color = "orange") +
  # point size & ranking
  geom_point(aes(size = n, color = Rank, alpha = 0.8)) +
  scale_color_continuous("Ranking", low = "#FFCC00", high = "#3300FF") +
  scale_size_continuous("Count", range = c(1, 12)) +
  labs(title = "Indoor TF girls (800m) - ranking & frequency") -> itf_girls_800m
```

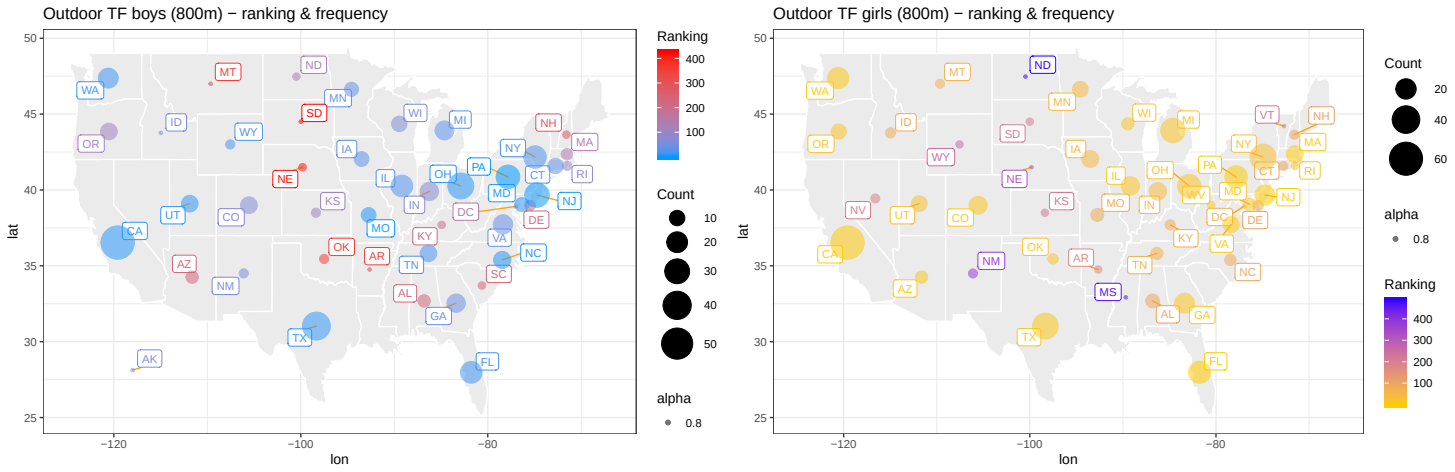
As we can see from the indoor results, the higher ranking athletes mostly from northeastern region.

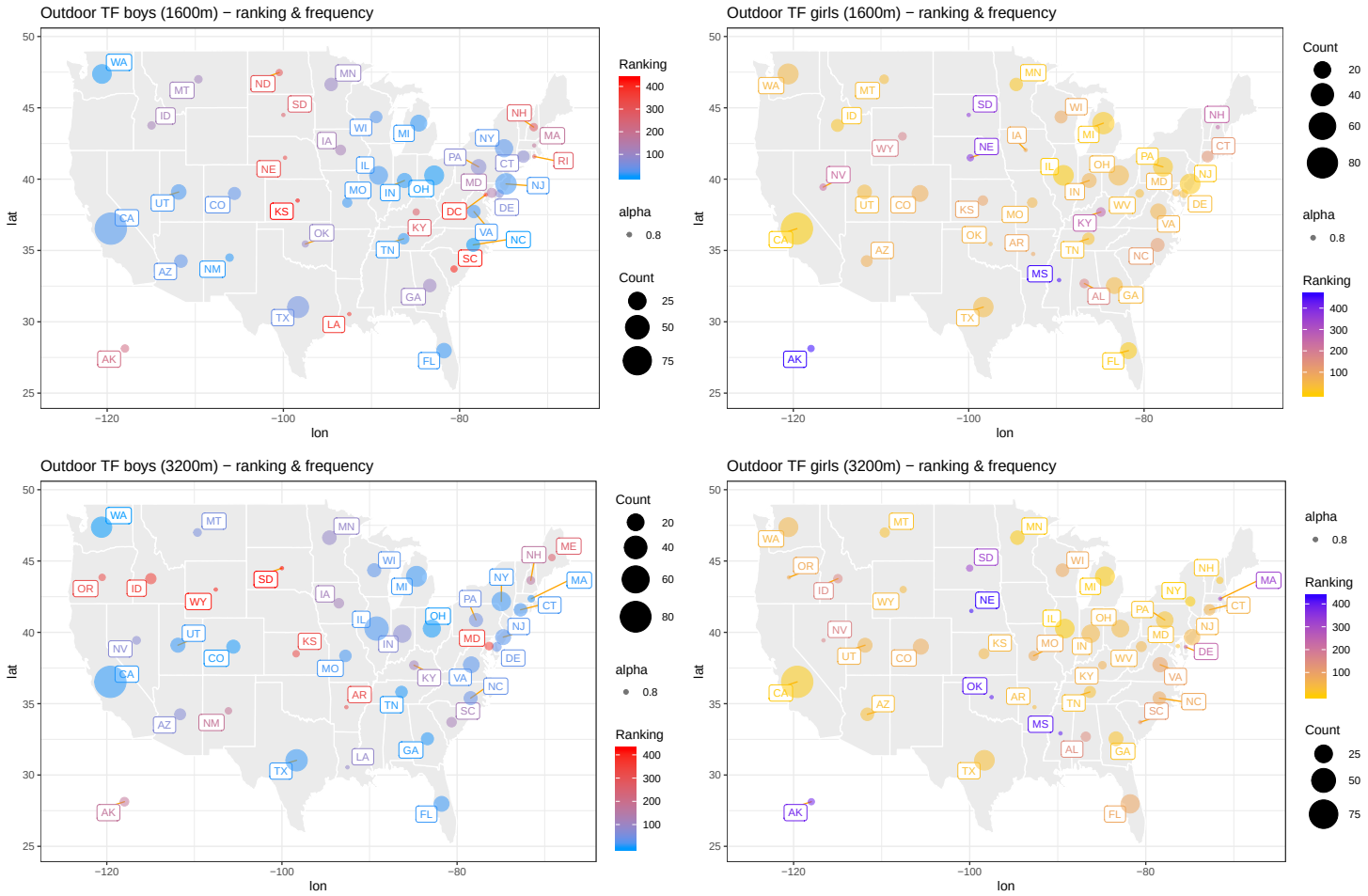




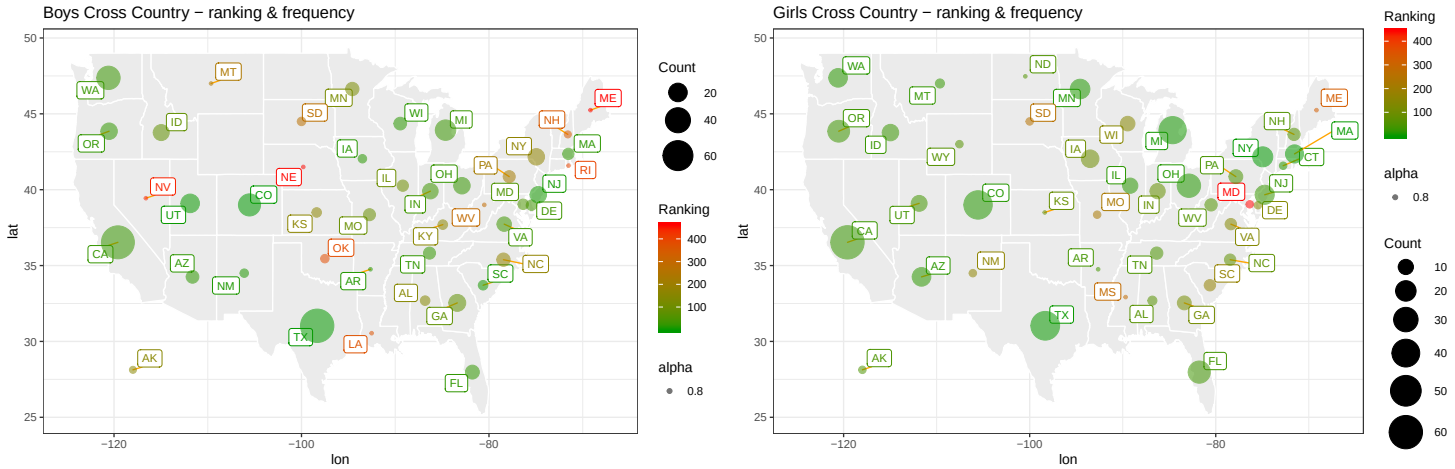
Since the codes are basically the same, I am hiding the rest of 1600m/3200m codes on the .pdf file.

- Outdoor Track & Field
  The outdoor results are clearly more spread out, but we can see the west coast states and the southern region states did a better job than the athletes who are from northeastern states.
  The codes are hidden on .pdf file





- Cross Country
  For the cross country data, I wouldn't not make a conclusion through this plot. Because the outstanding athletes are scatter around. It needs more research (pull up more years data, take a deeper look at their performance of time spending, etc.) to get a possible insight. *The codes are hidden on .pdf file*



- As we can see a similar statewide pattern between events in indoor/outdoor track & field data. I respectively combine the list of all indoor and outdoor datasets, and count the frequency by state (will use this data in Q2 and Q3).

```
# combine the list of datasets by indoor/outdoor
bind_rows(hs_indoor18) %>%
  group_by(state) %>%
```

```r
  count() %>%
  arrange(desc(n)) -> hs_in18_freq

plot_usmap(data = hs_in18_freq, values = "n", color = "white", labels = TRUE) +
  scale_fill_continuous(low = "#33CCFF", high = "#FF3300",
                        name = "# of athletes", label = scales::comma) +
  theme(legend.position = "right",
        panel.background = element_rect(colour = "Black")) +
  ggtitle("The distribution of Indoor Track & Field athletes") -> ITF_dist
```

The outdoor codes are hidden on .pdf file

Question 2

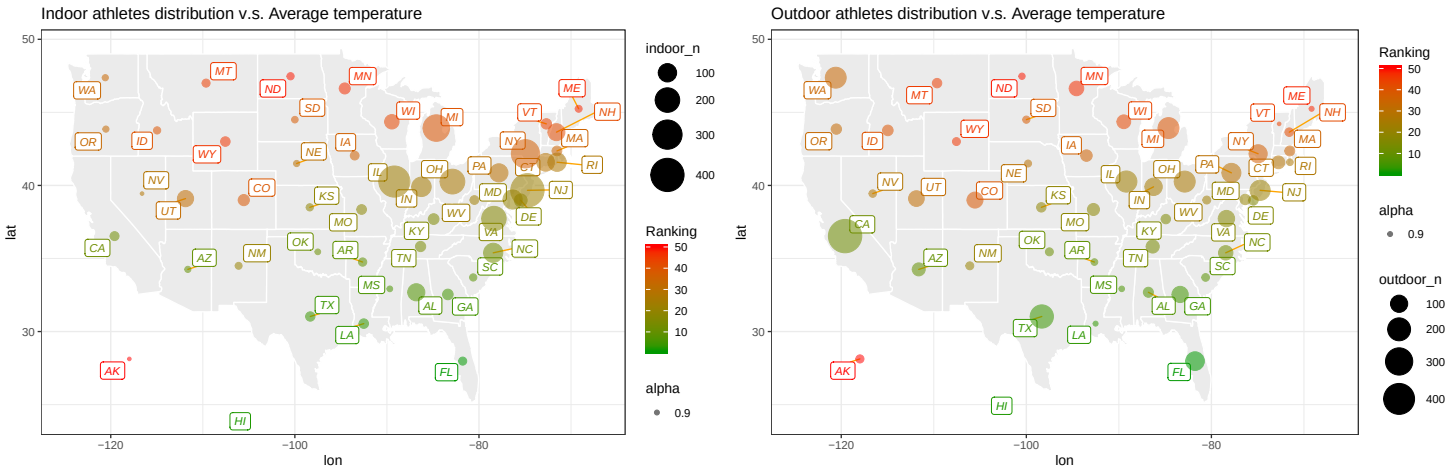
Does climate matter? How does the climate factors (temperature, rainfall inches, humidity, sunshine hours, wind speed, elevations) relate to elite athletes in the U.S.?

```
hs_freq <- hs_in18_freq %>%
  left_join(hs_out18_freq, by = "state") %>%
  set_names(c("state", "indoor_n", "outdoor_n"))
```

- Average Temperature: rank 1 means the warmest.
  It's cooler in the northern place, but the athletes' performances are not divided by north and south but east and west. I would suggest the average temperature is not significant on players average performance, but I believe the temperature will be crucial for each local competition.

```
# join data
temperature_hs <- temperature %>%
  left_join(hs_freq, by = "state")

# visualization
ggplot(temperature_hs, aes(x = lon, y = lat, color = Rank)) +
  geom_polygon(data = usmap, aes(x = long, y = lat, group = group),
               color = "white", fill = "grey92") +
  # label
  ggrepel::geom_label_repel(aes(label = state, fontface = "italic"),
                            data = temperature_hs, size = 3,
                            label.size = 0, segment.color = "orange") +
  # point size & ranking
  geom_point(aes(size = indoor_n, color = Rank, alpha = 0.9)) +
  scale_color_continuous("Ranking", low = "#009900", high = "red") +
  scale_size_continuous(range = c(1, 12)) +
  ggtitle("Indoor athletes distribution v.s. Average temperature") -> temp_ITF
```

```
ggarrange(temp_ITF, temp_OTF)
```

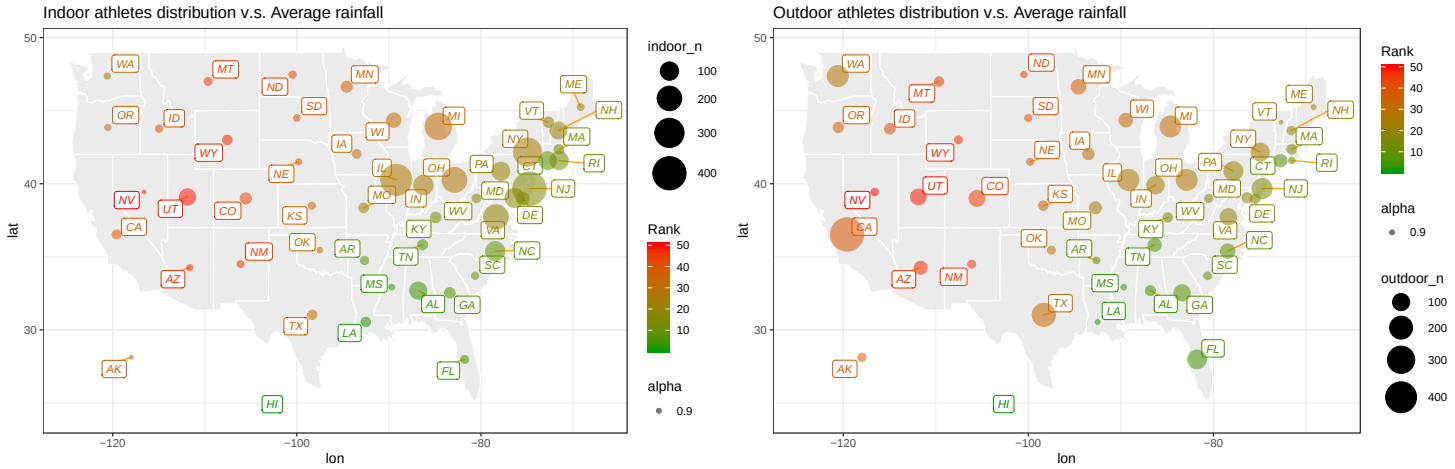


- Average Rainfall: rank 1 means more rain.
  The average rainfall is significant to the indoor/outdoor result, the eastern states tend to perform better at the indoor activities.

```
# join data
rainfall_hs <- rainfall %>%
```

```
  left_join(hs_freq, by = "state")

# visualization
ggplot(rainfall_hs, aes(x = lon, y = lat, color = Rank)) +
  geom_polygon(data = usmap, aes(x = long, y = lat, group = group),
               color = "white", fill = "grey92") +
  # label
  ggrepel::geom_label_repel(aes(label = state, fontface = "italic"),
                            data = rainfall_hs, size = 3,
                            label.size = 0, segment.color = "orange") +
  # point size & ranking
  geom_point(aes(size = indoor_n, color = Rank, alpha = 0.9)) +
  scale_color_continuous(low = "#009900", high = "red") +
  scale_size_continuous(range = c(1, 12)) +
  ggtitle("Indoor athletes distribution v.s. Average rainfall") -> rain_ITF
```

```
ggarrange(rain_ITF, rain_OTF)
```

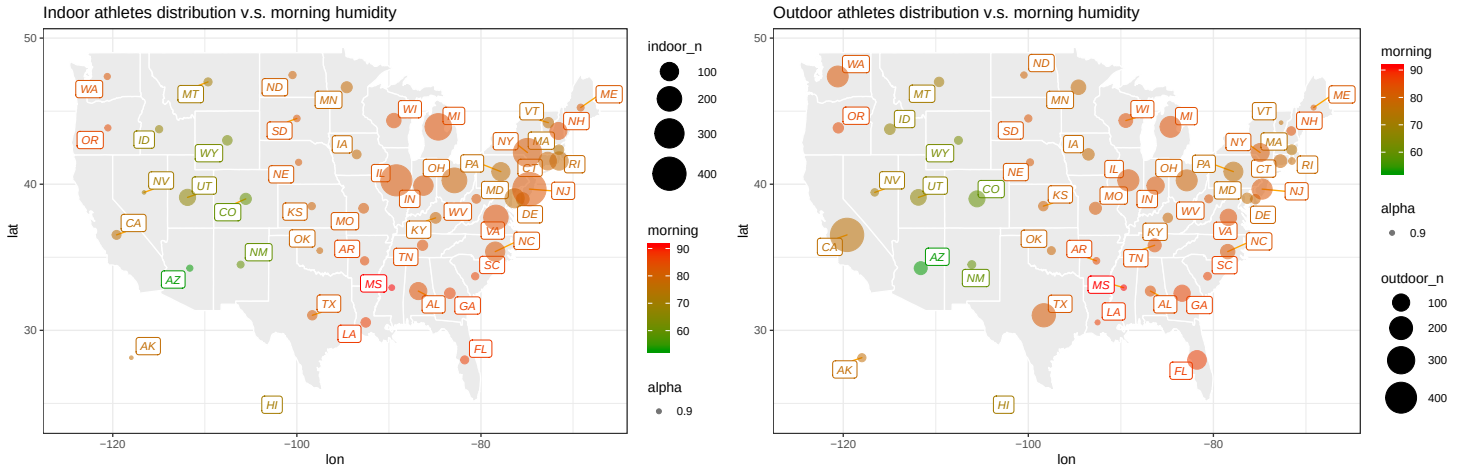


- Humidity: the humidity(%) in the morning usually comes with a higher humidity. The humidity is partially important as well.

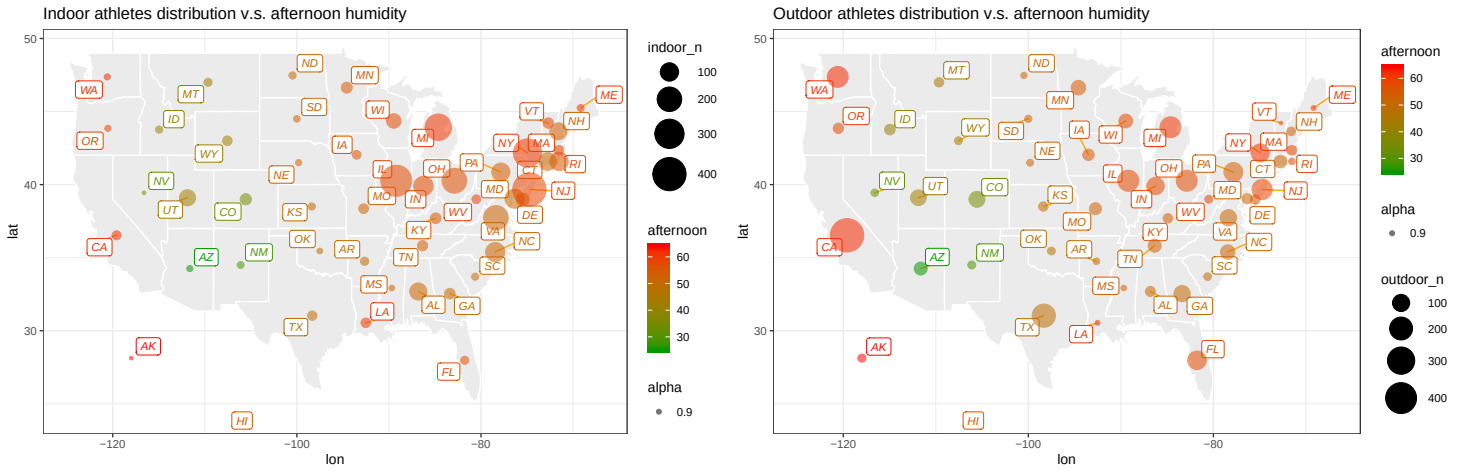
```
# join data
humidity_hs <- humidity %>%
  left_join(hs_freq, by = "state")

# --------- Morning ---------
ggplot(humidity_hs, aes(x = lon, y = lat, color = morning)) +
  geom_polygon(data = usmap, aes(x = long, y = lat, group = group),
               color = "white", fill = "grey92") +
  # label
  ggrepel::geom_label_repel(aes(label = state, fontface = "italic"),
                            data = humidity_hs, size = 3,
                            label.size = 0, segment.color = "orange") +
  # point size & ranking
  geom_point(aes(size = indoor_n, color = morning, alpha = 0.9)) +
  scale_color_continuous(low = "#009900", high = "red") +
  scale_size_continuous(range = c(1, 12)) +
  ggtitle("Indoor athletes distribution v.s. morning humidity") -> morn_humid_ITF
```

- Showing all plots (morning/afternoon indoor/outdoor)

```
ggarrange(morn_humid_ITF, morn_humid_OTF)
```




```
ggarrange(aft_humid_ITF, aft_humid_OTF)
```

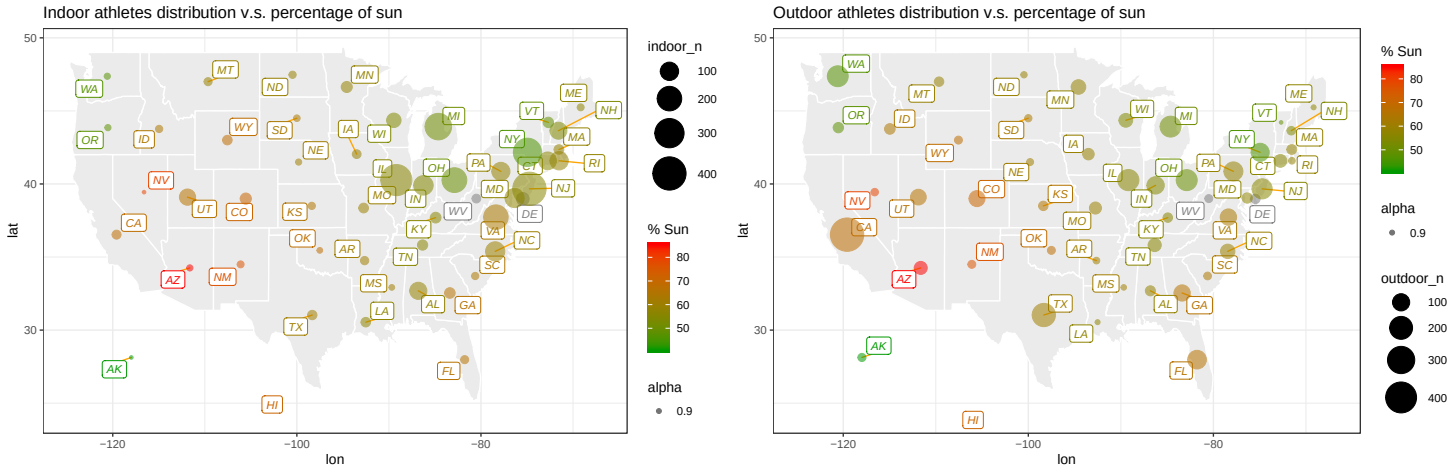


- Sunshine hours
  The more sunshine hours the state has, the more better outdoor athletes they would have.

```
# join data
sunshine_hs <- sunshine %>%
  left_join(hs_freq, by = "state")

# visualization
ggplot(sunshine_hs, aes(x = lon, y = lat, color = `% Sun`)) +
  geom_polygon(data = usmap, aes(x = long, y = lat, group = group),
               color = "white", fill = "grey92") +
  # label
  ggrepel::geom_label_repel(aes(label = state, fontface = "italic"),
                            data = sunshine_hs, size = 3,
                            label.size = 0, segment.color = "orange") +
  # point size & ranking
  geom_point(aes(size = indoor_n, color = `% Sun`, alpha = 0.9)) +
  scale_color_continuous(low = "#009900", high = "red") +
```

```
  scale_size_continuous(range = c(1, 12)) +
  ggtitle("Indoor athletes distribution v.s. percentage of sun") -> sun_ITF
```

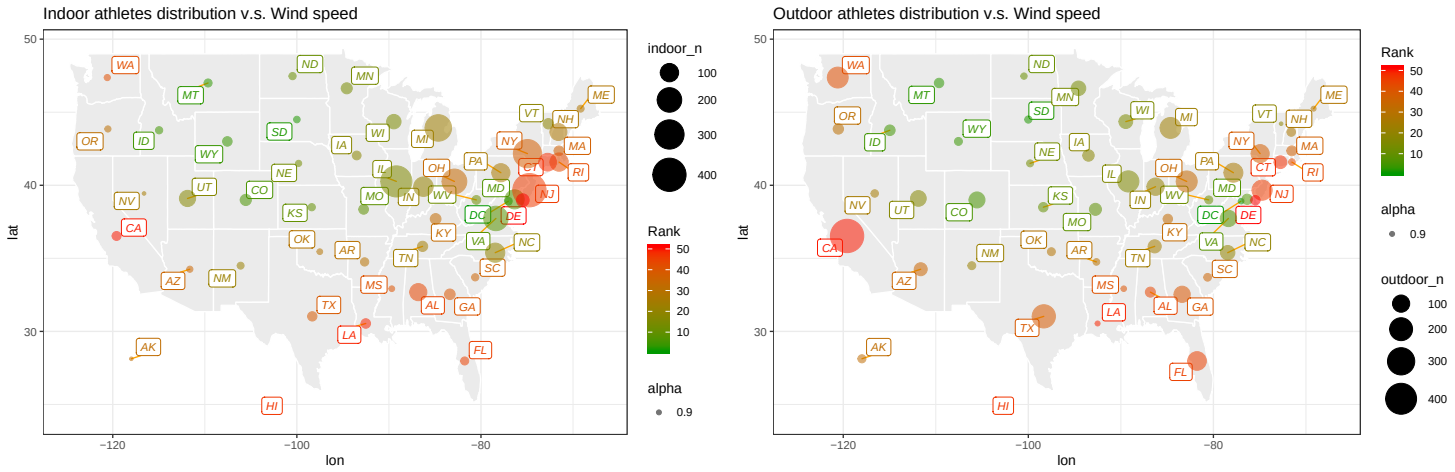
```
ggarrange(sun_ITF, sun_OTF)
```



- Windspeed: rank 1 suggests a stronger wind speed.
  The stronger wind speed by inland north America suggests less well performanced athletes.

```
# join data
windspeed_hs <- windspeed %>%
  left_join(hs_freq, by = "state")

# visualization
ggplot(windspeed_hs, aes(x = lon, y = lat, color = Rank)) +
  geom_polygon(data = usmap, aes(x = long, y = lat, group = group),
               color = "white", fill = "grey92") +
  # label
  ggrepel::geom_label_repel(aes(label = state, fontface = "italic"),
                            data = windspeed_hs, size = 3,
                            label.size = 0, segment.color = "orange") +
  # point size & ranking
  geom_point(aes(size = indoor_n, color = Rank, alpha = 0.9)) +
  scale_color_continuous(low = "#009900", high = "red") +
  scale_size_continuous(range = c(1, 12)) +
  ggtitle("Indoor athletes distribution v.s. Wind speed") -> windsp_ITF
```

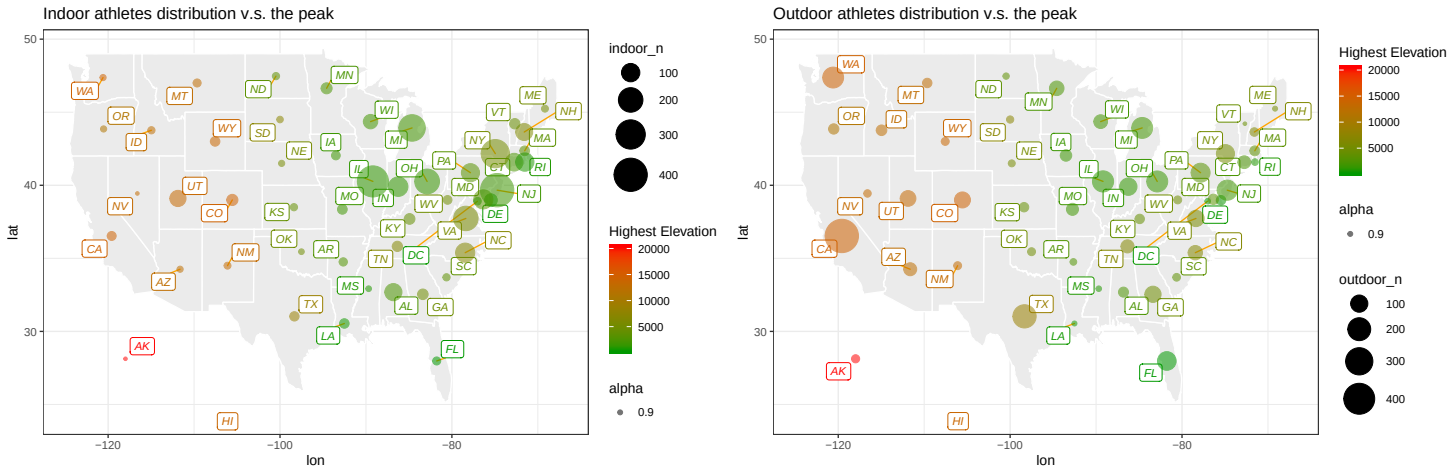
```
ggarrange(windsp_ITF, windsp_OTF)
```



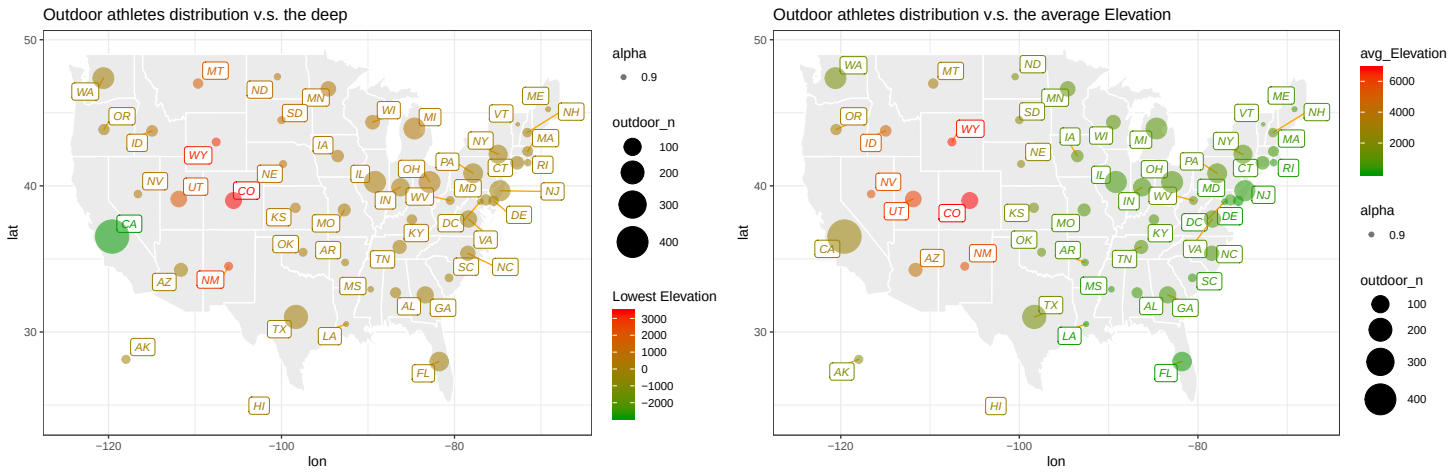
- Elevation
  According to the upper two plots, the western region in the North America tends to have a higher highest elevation than in the eastern. I found the indoor list is unrelated to the elevation. However, the last plot with average elevation versus outdoor performance list, the middle west side has higher average elevation, then the west coast is in the second place, the lowest average elevation is in the east region. But the performance list gives a converse result, people in the west (higher elevation) generally have a better performance than the east region (lower elevation) for the outdoor competitions.

```
# join data
elevation_hs <- elevation %>%
  left_join(hs_freq, by = "state")

# visualization
ggplot(elevation_hs, aes(x = lon, y = lat, color = `Highest Elevation`)) +
  geom_polygon(data = usmap, aes(x = long, y = lat, group = group),
               color = "white", fill = "grey92") +
  # label
  ggrepel::geom_label_repel(aes(label = state, fontface = "italic"),
                            data = elevation_hs, size = 3,
                            label.size = 0, segment.color = "orange") +
  # point size & ranking
  geom_point(aes(size = indoor_n, color = `Highest Elevation`, alpha = 0.9)) +
  scale_color_continuous(low = "#009900", high = "red") +
  scale_size_continuous(range = c(1, 12)) +
  ggtitle("Indoor athletes distribution v.s. the peak") -> elev_peak_ITF
```

```
ggarrange(elev_peak_ITF, elev_peak_OTF)
```



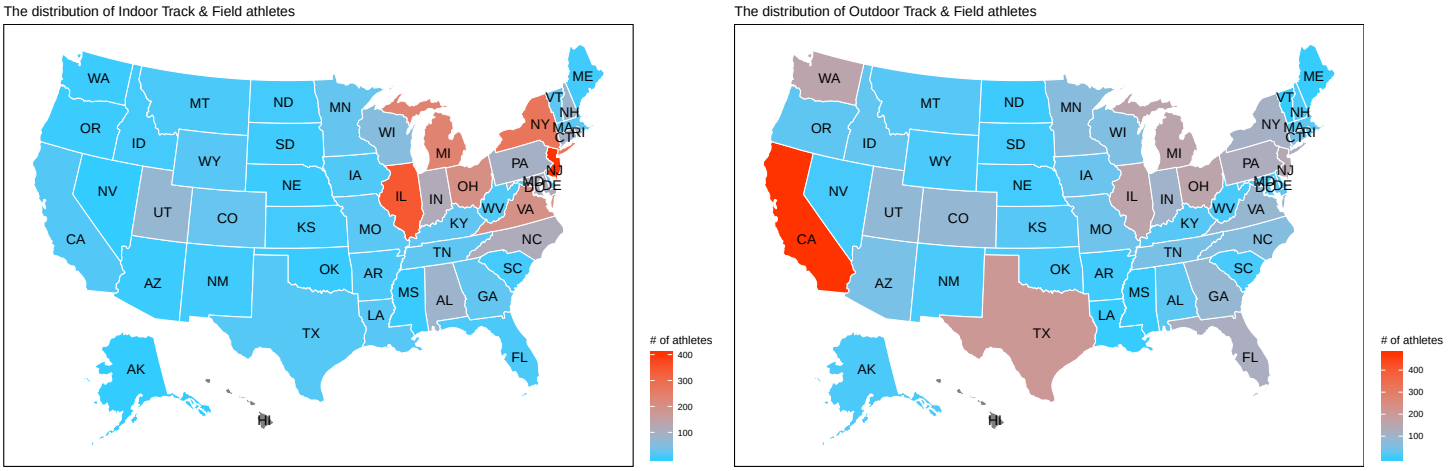

```
ggarrange(elev_deep_OTF, elev_avg_OTF)
```



Question 3

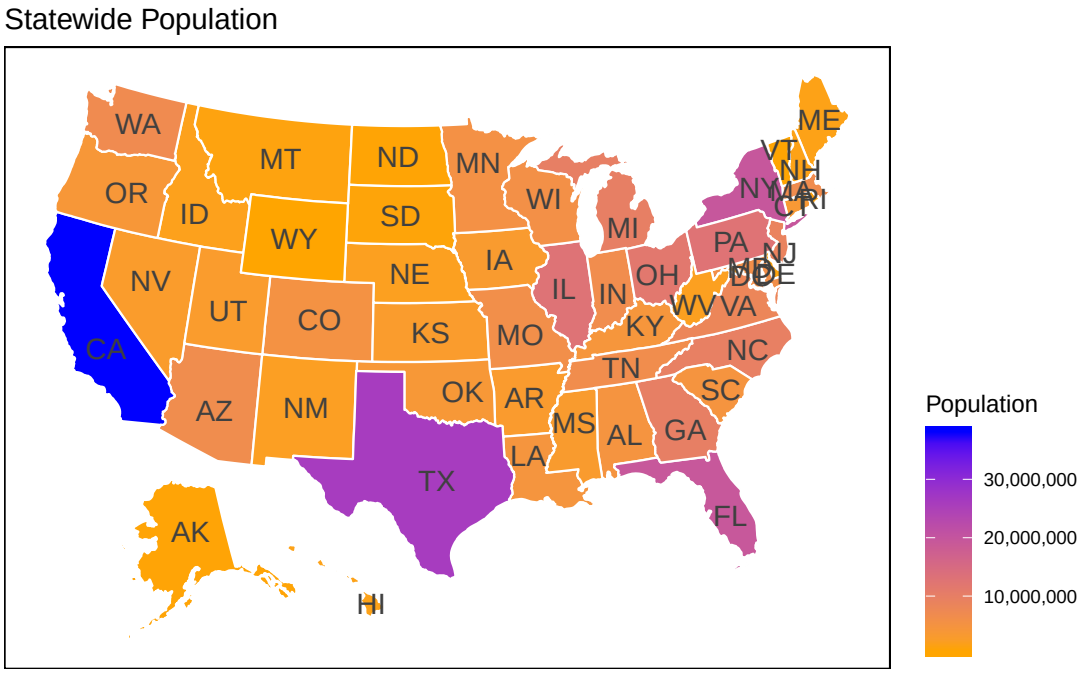
`How wealthy are those outstanding athletes? Is there any lower-income state that tends to have a better performance of athletes?`

- The distribution of elite athletes by indoor and outdoor Track & Field competitions. This map is similar to the household income. Where with more population, in which produces more outstanding athletes.



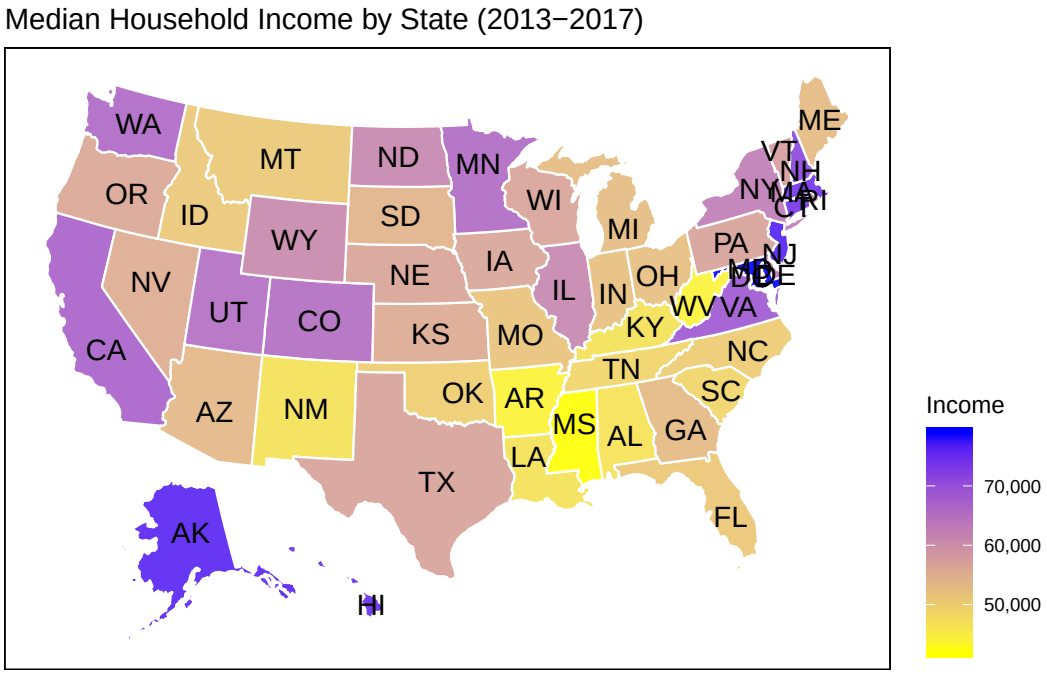
- Statewide Population: the population map is similar to the player performance as well as median income.

```
plot_usmap(data = windspeed, values = "Population", color = "white",
           labels = TRUE, label_color = "grey25") +
  scale_fill_continuous(low = "orange", high = "blue",
                        name = "Population", label = scales::comma,
                        limits = c(575250, 38066921)) +
  theme(legend.position = "right",
        panel.background = element_rect(colour = "Black")) +
  labs(title = "Statewide Population")
```



- Median House hole Income: we found there are more population in Texas, Florida, Ohio, and Pennsylvania. The athletes from Texas and Florida are good at outdoor Track & Field competitions (see the distribution of outdoor TF), and the athletes from Ohio and Pennsylvania are better at indoor Track & Field games. What's more, the median household income from these 4 states are lower. It's suggestible to recruit possible well performance and middle class athletes in these states.

```
plot_usmap(data = median_income,
           values = "Income", color = "white", labels = TRUE, alpha = 0.9) +
  scale_fill_continuous(low = "#FFFF00", high = "#0000FF",
                        name = "Income", label = scales::comma) +
  theme(legend.position = "right",
        panel.background = element_rect(colour = "Black")) +
  labs(title = "Median Household Income by State (2013-2017)")
```



Conclusion & Future Works

In summary, people who engage in Track & Field and Cross Country are mostly from popular states such as Washington, California, Colorado, Utah, Texas, Florida, Massachusetts, New Jersey, and New York. Besides, these states also have a larger population and higher household income than other states, which means that they have a bigger size of statistical population so that those states accordingly produce more well performanced athletes. Nevertheless, one interesting results I found is that athletes who are from the northeastern United States generally performanced well in `indoor` Track & Field games; however, western, west coast and southern of the United State's athletes did a better job in `outdoor` Track & Field competitions. One possible reason to explain this phenomenon is the climate condition. The climate data results suggested that players' indoor/outdoor performance relates to natural conditions. Which makes sense that the northeastern U.S. is colder, more humid, and less sunshine hours, people would prefer indoor training; in contrast, the west coast and southern of the United States are drier, hotter, less rain, and longer sunshine hours, so in effect the outdoor training is easy and preferable.

On the other hand, elite female athletes' are more spread out, they are from many different states instead of gathering between popular states like male athletes. As a matter of fact, the athletes' performance list, climate data by states, and the the population are interrelating. A better climate condition appeals more people to live, a state with more residents gives a larger population size, and a bigger amount of people

tends to be more competitive and outstanding. Nevertheless, as we know the respective features of western and northeastern athletes, one may conclude that if the coach would like to recruit indoor Track & Field professionals. The residents from Pennsylvania and Ohio may have some well performance but middle class athletes. For outdoor Track & Field athletes, people from Florida tend to earn a lower income but the amount of outstanding athletes in Florida is actually good enough to beat the athletes in California. Last but not least, players from Texas will be a good second choice.
Although there are so many research studies indicate the climate and environment is significant to athletes' performance, my data in fact didn't bother with climate data that much. One possible reason might be athletes nowadays would rather to train themselves in the environment where they can do well. Athletes from Northeastern region usually work on indoor training, and athletes from warmer places focus on the outdoor Track & Fields. Both side basically find their expertise and partially reduce the climate factors. However, I believe the climate factors will be interesting if we analyze their performance between western and eastern outdoor Track & Field competitions. It can be a suggestible future work. This preliminary study helps us to find a insight and a possible phenomena among U.S. high school athletes. As we do find some interesting results between the athletes. It would be helpful to do some further analysis, such as pull up the data through athlete's hometown versus the climate condition between countries, or make it deeper for exploring the yearly changes.

The URL for Shiny App: https://wafer110.shinyapps.io/Shiny_WeiHuaHsu/

Reference

- NOAA (National Oceanic and Atmospheric Administration)
  httpd://WWW.CDC.nova.gov/ca/statewide/mapping/110/PCP/201812/12/value

- Current Results
  ://WWW.current results.com/Weather/US/average-annual-state-sunshine.pp
  https://www.currentresults.com/Weather/US/annual-average-humidity-by-state.php

- USA.com
  http://www.usa.com/rank/us--average-wind-speed--state-rank.htm

- Smart Search
  https://smart-search.app/resources/2019-2020-efc-quick-reference.pdf

- Census Bureau
  https://www.census.gov/search-results.html?searchType=web&cssp=SERP&q=annual%20income

- infoplease.com
  https://www.infoplease.com/world/united-states-geography/highest-lowest-and-mean-elevations-united-state

- MileSplit
  https://www.milesplit.com/

- TFRRS
  https://tfrrs.org/

- American University Athletics
  https://aueagles.com/sports/track-and-field/roster#sidearm-m-roster