Project 1

```
olympics <- readr::read_csv('https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/data/

triathlon <- olympics %>%
  filter(!is.na(height)) %>%              # only keep athletes with known height
  filter(sport == "Triathlon") %>%        # keep only triathletes
  mutate(
    medalist = case_when(                 # add column to track medalist vs not
      is.na(medal) ~ "non-medalist",
      !is.na(medal) ~ "medalist"          # any medals (Gold, Silver, Bronze) count
    )
  )
```

`triathlon` is a subset of `olympics` and contains only the data for triathletes. More information about the original `olympics` dataset can be found at https://github.com/rfordatascience/tidytuesday/tree/master/data/2021/2021-07-27/readme.md and https://www.sports-reference.com/olympics.html.

```
triathlon
```

```
## # A tibble: 528 x 16
##       id name     sex     age height weight team  noc   games  year season city
##    <dbl> <chr>    <chr> <dbl>  <dbl>  <dbl> <chr> <chr> <chr> <dbl> <chr>  <chr>
##  1   579 "Anasta~ F        26    164     51 Russ~ RUS   2016~  2016 Summer Rio ~
##  2   623 "Irina ~ F        27    175     56 Russ~ RUS   2008~  2008 Summer Beij~
##  3   623 "Irina ~ F        31    175     56 Russ~ RUS   2012~  2012 Summer Lond~
##  4   748 "Mariko~ F        29    162     51 Japan JPN   2012~  2012 Summer Lond~
##  5  1190 "Simon ~ M        31    179     64 Aust~ AUT   2008~  2008 Summer Beij~
##  6  2271 "Fernan~ M        25    178     67 Spain ESP   2016~  2016 Summer Rio ~
##  7  2331 "Marko ~ M        25    189     78 Esto~ EST   2004~  2004 Summer Athi~
##  8  2331 "Marko ~ M        29    189     78 Esto~ EST   2008~  2008 Summer Beij~
##  9  2849 "Kather~ F        34    163     53 Aust~ AUT   2004~  2004 Summer Athi~
## 10  2849 "Kather~ F        38    163     53 Aust~ AUT   2008~  2008 Summer Beij~
## # i 518 more rows
## # i 4 more variables: sport <chr>, event <chr>, medal <chr>, medalist <chr>
```

Introduction: This project involves working with the `triathlon` dataset, which comprises 528 triathletes' records from triathlon events in the Olympics. Each row in this dataset contains information about one athlete, encompassing 16 columns that provide details about demographic, participation, and performance information of the triathletes. Demographic information covers triathletes' name, sex, age, height, and weight. Participation details includes the athlete's team, NOC (National Olympic Committee) region, Olympic Games name, year, season, host city, and the specific event in which the athlete competed. The performance of each athlete is reflected in the type of medal awarded and whether or not they achieved medalist status.

Five variables were used to answer the three questions above: the triathletes' country or team of competition (column `team`), gender (column `sex`), the year of the Olympics in which they competed (column `year`), their height (column `height`), and whether or not they achieved medalist status (column `medalist`). The team represents the Country name; Gender is encoded as F/M, where F stands for female and M stands for male. The year of the Olympics is provided in years, and height is represented in centimeters. Each athlete's medalist status is categorized as either non-medalist or medalist.

Approach: To determine the number of events male and female triathletes have competed for each country, we will use bar charts (`geom_bar`) and stack them to distinguish each gender Bar charts will help us visualize a comparison of gender-specific participation across countries.

Using box plots (`geom_box`), we will compare the height distributions between male and female triathletes, and investigate height trends over time by considering the `year` column.

Finally, we will examine height variations among triathletes, considering both medal status ('medalist') and sex ('sex') using sinaplot (`geom_sina`) overlayed over violin plot (`geom_violin`). This combined approach provides an extra layer of detail, enhancing our understanding of height patterns within these distinct groups.

Analysis:

1. In how many events total did male and female triathletes compete for each country?

To address this question, we will create a box plot to visualize the frequencies of events in which each country competed by gender.

```
#This code creates a bar chart that counts the total events each country participated in,
#categorizing them by sex, and uses facet_wrap to separate out each country. It is in
#descending order.

ggplot(triathlon, aes(x ="", fill = sex)) +

  # Add bars with specified width and dodge position
  geom_bar(
    width = 0.75,
    position= "dodge"
  )+
  # Set color palette for sex
  scale_fill_brewer(
    name = "Sex",
    palette = "Set1",
    labels = c("Female", "Male")
  )+

  # Customize x-axis title
  scale_x_discrete (
    name = 'Country'
  )+

  #Customize y-axis title and breaks
  scale_y_continuous(
    name = 'Total Events',
    breaks = seq(1, 15, by = 2),
    expand = expansion(mult = c(0, 0.3))
  )+

  #Add label for each bar that represents the count
  geom_text(
    aes(label = after_stat(count)),
    stat = "count",
    vjust = -0.2,
    colour = "black",
    position = position_dodge(.9)
  )+
  # Facet wrap by country in descending order.
```

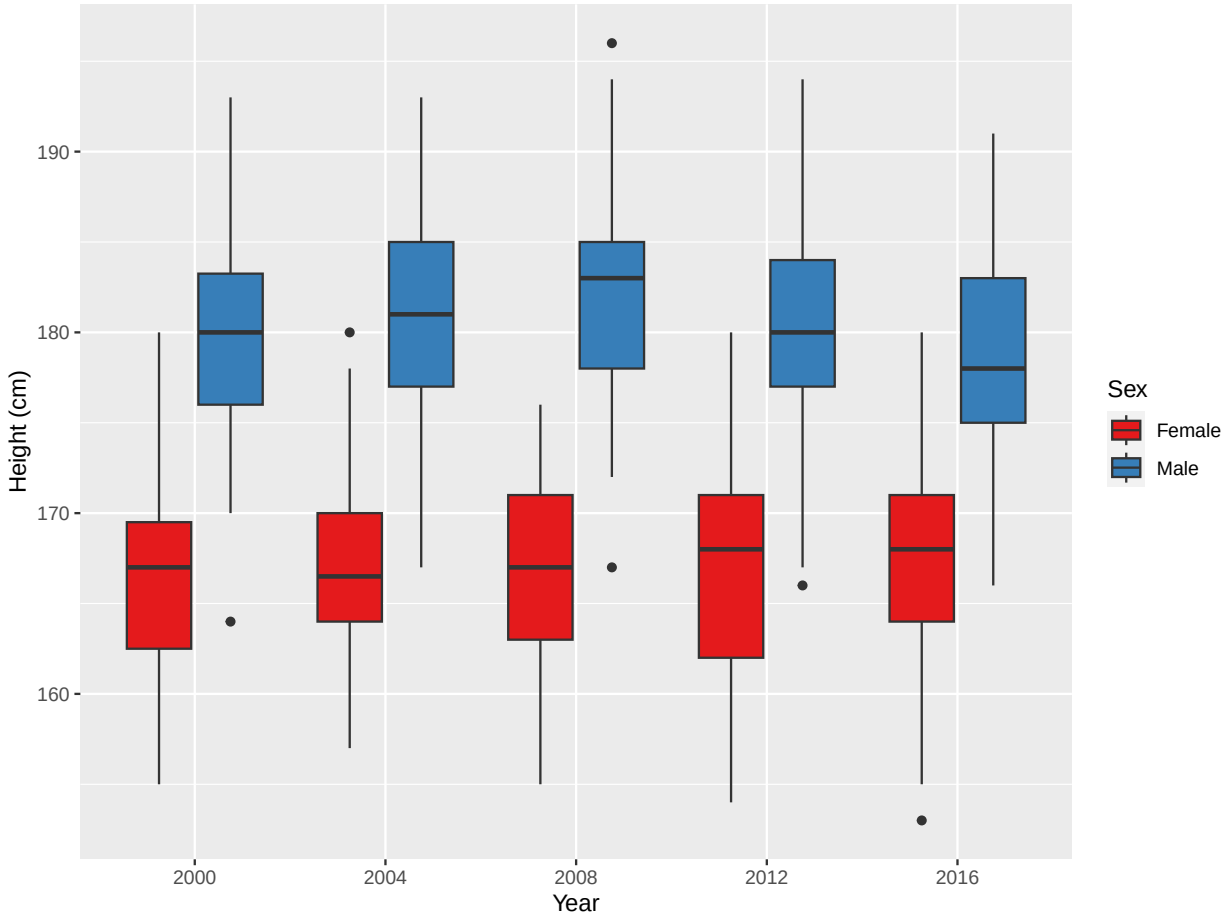
```
  facet_wrap(
    ~fct_infreq(team)
  )
```

The highest number of events, totaling twenty-nine, was observed for Australia, the USA, and Great Britain. Australia and the United States had one more female participant than Great Britain. In contrast, Syria and South Korea, among eleven others, had the lowest participation with just one event, and the majority of the participants in that event were male.

2. Are there height differences among triathletes between sexes or over time? To see the differences among triathletes between sexes or over time, we build a box plot.

```
#This code creates a box plot that shows the distribution of athlete's height across year and
#categorize them into sex.
```

```r
ggplot(triathlon, aes(x = as.factor(year), y = height, fill = sex)) +

  #Build box plot
  geom_boxplot() +

  #Label the sex and set the color of each category
  scale_fill_brewer(
    name = "Sex",
    labels = c("Female", "Male"),
    palette = "Set1"
    )+

  #Label the title of the x axis
  scale_x_discrete (
    name = 'Year'
  ) +

  #Label the title of the y-axis
  scale_y_continuous(
    name = 'Height (cm)'
  )
```



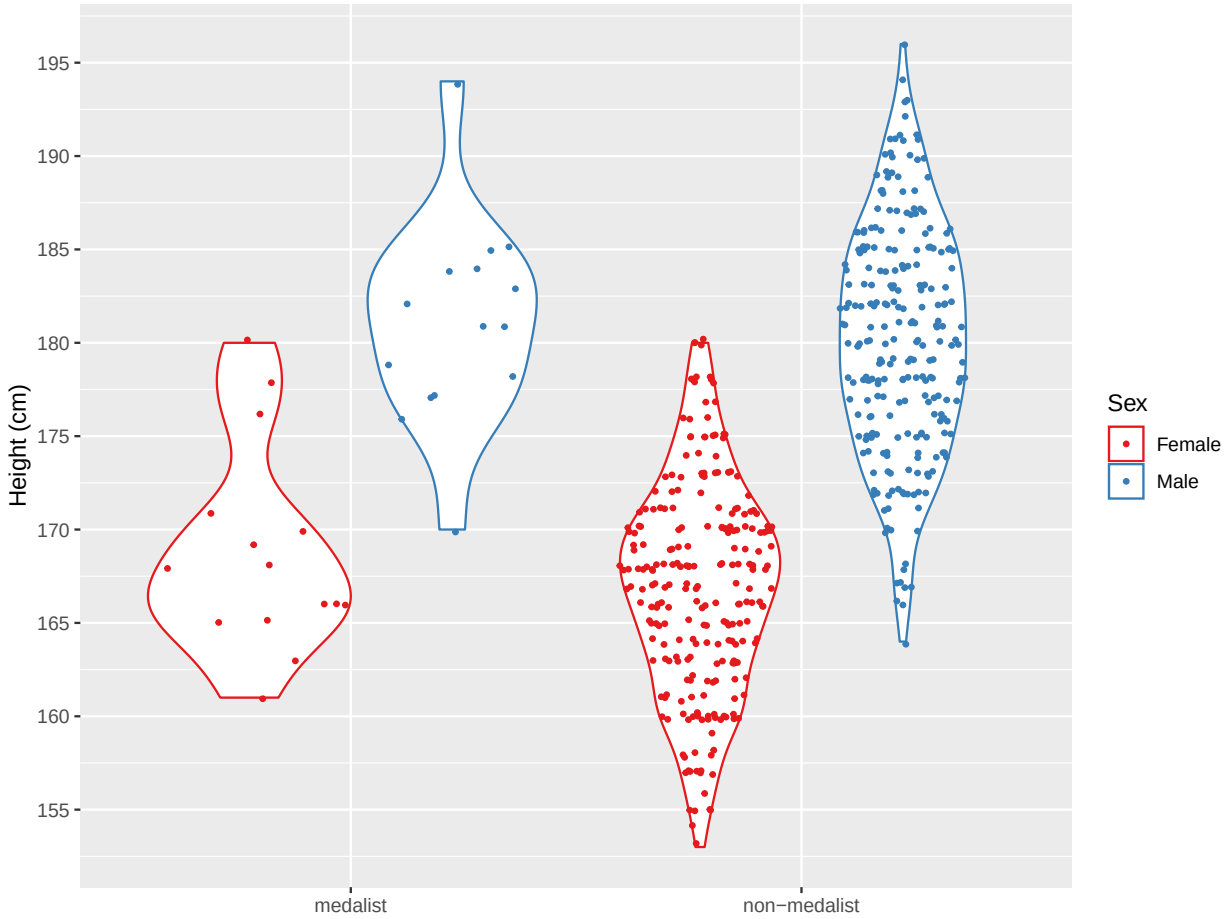
The average height of male athletes typically centers around 180cm, though exceptions were noted in 2008 and 2015. The median height experienced slight variations during these years, reflecting both higher and

lower measurements. The distribution of heights stay fairly constant with the inter quartile range staying in the range of 175cm to 185cm. For female athletes, the inter quartile range is around 163cm to around 170cm, with the central tendency around 167cm. Remarkably, these height distributions have demonstrated consistent patterns over the years.

3. Are there height differences among triathletes that have medaled or not, again also considering athlete sex? **Hints:**

To answer this question, we will employ a visualization strategy combining sina plots overlaid on violin plots. We utilize this type of plot because as sina plots offer a detailed representation of height density and distribution, surpassing the limitations of traditional box plots. The overlaid violin plot complements this by providing a holistic view of the overall shape of height distribution, differentiated by both gender and medalist status.

```
#This code creates a sina plot overlayed over a violin plot that shows the distribution of height by me
#categorized by sex.

ggplot(triathlon, aes(x = medalist, y = height, color = sex)) +

  #Create violin plot
  geom_violin() +

  #Create sina plot
  geom_sina(size = 0.8) +

  #Label the legend and set the color of sex category
  scale_color_brewer(
    name = "Sex",
    labels = c("Female", "Male"),
    palette="Set1"
    )+

  #Don't label the x axis
  scale_x_discrete(
    name = NULL
  ) +

  #Label the y-axis and give them breaks ranging from 150cm to 200cm with steps of 5cm
  scale_y_continuous(
    name = "Height (cm)",
    breaks = seq(150, 200, by = 5),
  )
```



Medalist heights exhibit a range of approximately 160cm to 180cm for females and 170cm to just over 190cm for males. In contrast, non-medalist heights vary from about 153cm to 180cm for females and approximately 163cm to slightly over 195cm for males.

Discussion: The observation of event participation across countries in the triathlete dataset reveals that the highest number of events, totaling twenty-nine, was concentrated in Australia, the USA, and Great Britain, showcasing their significant involvement in triathlon competitions. Conversely,Syria and South Korea, along with eleven others, exhibited a minimal presence, each hosting only one event. A more detailed examination of the dataset reveals that over half of the countries had fewer than 11 participants. The observed pattern, where more developed European countries dominated the higher end of participation, while less developed nations had comparatively fewer participants, suggests a possible correlation between triathlon prominence and the developmental status of countries.

Both female and male athletes' median height unveils slight variations across different years, reflecting both higher and lower measurements. Despite these fluctuations, the overall distribution of heights remains relatively constant. The observed patterns in male and female height distributions reflect a common physical characteristics within the sport of triathlon; this stability of height trends over the years suggests that athletes possess a relatively uniform height that aligns with the demands of the sport.

Medalist heights exhibit a more concentrated and wider bell shape, suggesting a focal range of heights. However, it's we can consider that this characteristic might be influenced by the potentially smaller sample size of medalists. In contrast, non-medalist heights display an elongated and narrower shape, indicating a more diverse distribution across a broader range of heights. The pronounced differences are further highlighted by the larger min and max distributions for non-medalists, hinting at a heightened variability in height within this group. While the plot shows a prominent pattern at specific heights for medalist, we have to cautious in

concluding that height is related to winning a medal.