Project 1

We will work with the dataset `olympics_top` that contains data for the Olympic Games from Athens 1896 to Rio 2016 and has been derived from the `olympics` dataset. More information about the dataset can be found at: https://github.com/rfordatascience/tidytuesday/blob/master/data/2021/2021-07-27/readme.md The dataset, `olympics_top`, contains four new columns: `decade` (the decade during which the Olympics took place), `gold` (whether or not the athlete won a gold medal), `medalist` (whether or not the athlete won any medal) and `medal` (if the athlete won "Gold", "Silver", "Bronze" or received "no medal").

Part 1

Question: Which sports have the tallest or shortest athletes? And does the distribution of heights change for the various sports between medalists and non-medalists?

We recommend you use box plots for the first part of the question and use a ridgeline plot for the second part of the question.

Hints:

- To order boxplots by the median, you may have add the following to your ordering function to remove missing values before ordering: `na.rm = TRUE`
- To trim the tails in your ridgeline plot, you can set `rel_min_height = 0.01` inside `geom_density_ridges()`.

NAME: Vrishank Jannu, EID: vgj95

Introduction:

`olympics_top` is the data-set we will be working with which contains Olympic Games data from the Athens 1896 games all the way up to Rio 2016. Each record contains essential information of the athlete including their `name`, `age`, `sex`, `height`, nationality (`noc`), the `games` at which they played, `season`, `year`, `decade`, location of the olympic games (`city`), `event`, and `sport`. Additionally, each record contains information on whether or not the athlete won a gold medal (`gold`), whether or not the athlete won any medal (`medalist`) and if the athlete won a gold, silver, bronze, or no medal (`medal`).

To answer the question in Part 1 of which sports have the tallest or shortest athletes and if the distribution of heights change for various sports between the medalists and non-medalists, we will work with three variables, the athlete's `sport` and `height` and whether or not the athlete won a medal ('medalist').

Approach:

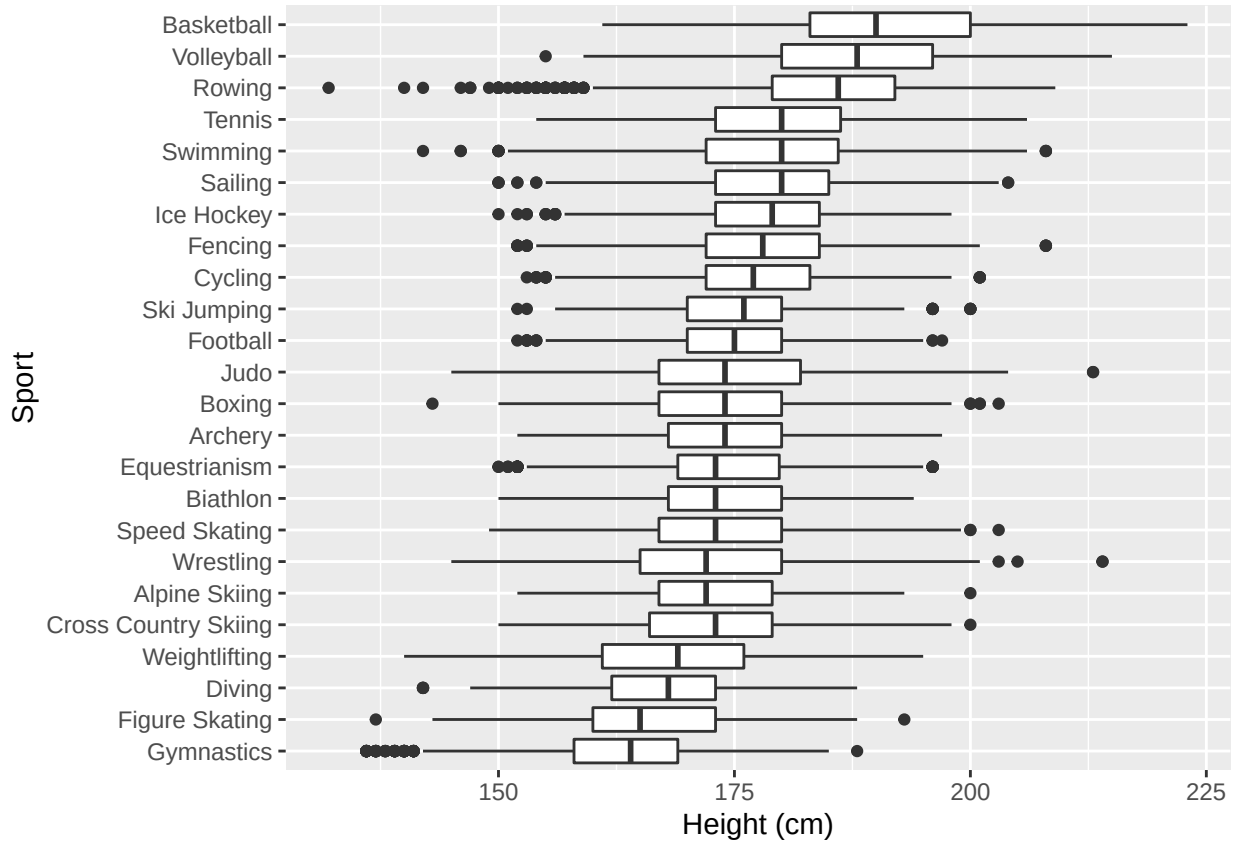
Our approach is to use a boxplot (`geom_boxplot()`) to answer the first part of the question in Part 1 which will show the height distributions of each sport. Boxplots are useful when we want to visualize many distributions at once. In this case, we want to visualize 24 different distributions (24 sports total). Boxplots are especially useful for clearly identifying the median of a distribution.

One limitation of boxplots is that it hides multimodality and other important features of distributions. In our case, we want to separate the medalists and non-medalists in each distribution to see if the distribution of heights in each sport change based on athletes who earn medals or not, which cannot be achieved visually with a boxplot. We will use a ridgeline plot (`geom_density_ridges()`) to introduce our third variable `medalist` and achieve multimodality.

Analysis:

```
ggplot(olympics_top, aes(x=reorder(sport, height, na.rm = TRUE), y=height)) +
  geom_boxplot() + coord_flip() + xlab("Sport") + ylab("Height (cm)")
```

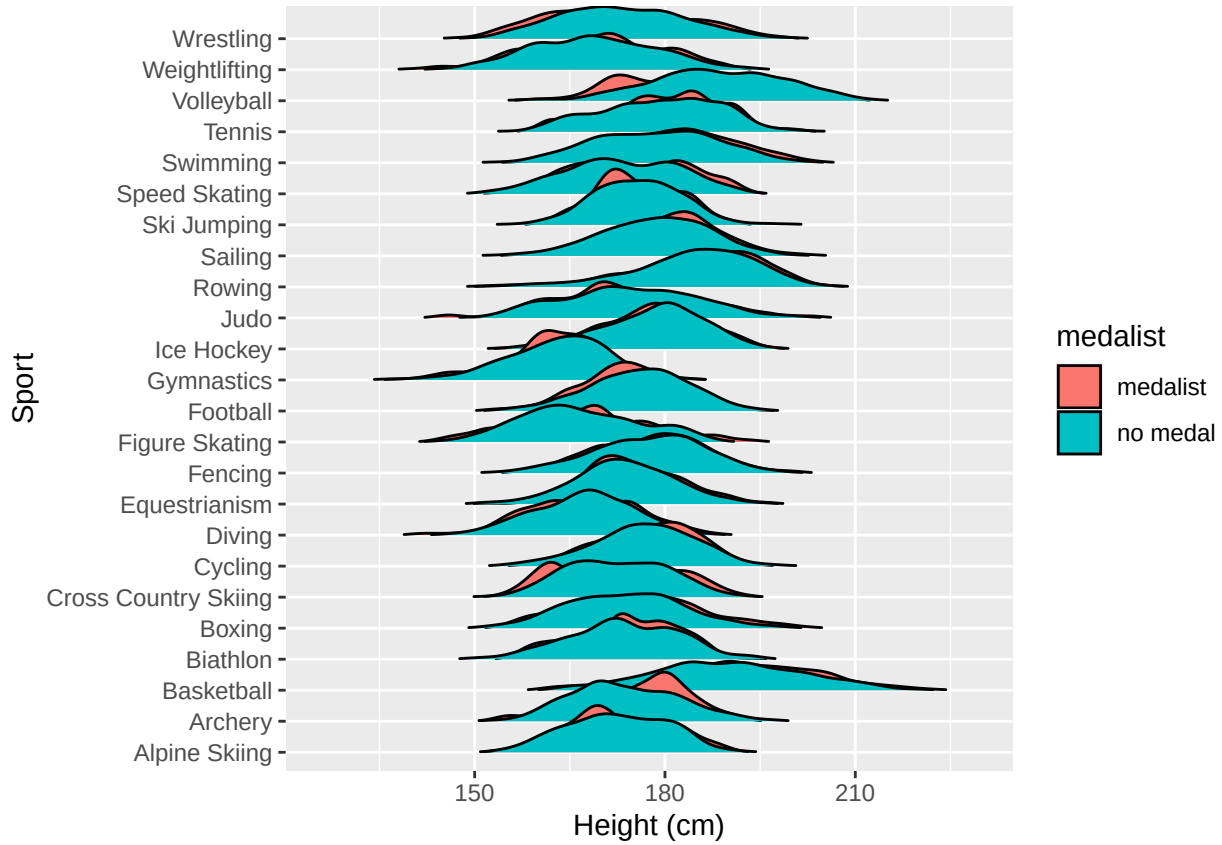
```
## Warning: Removed 19103 rows containing non-finite values (stat_boxplot).
```




```
ggplot(olympics_top,
       aes(height, sport, fill = medalist)) +
  geom_density_ridges(rel_min_height = 0.01) + xlab("Height (cm)") + ylab("Sport")
```

```
## Picking joint bandwidth of 2.19
```

```
## Warning: Removed 19103 rows containing non-finite values (stat_density_ridges).
```



Discussion: *From both plots, it is evident that the sport of basketball contains the tallest athletes. Volleyball and rowing are second and third on the list respectively. We can can conclude this based on the three largest median heights shown in the boxplots as well as the three rightmost ridges in the ridgelines distribution. In the sport of basketball, having a height advantage against other players allows you to create shots, dunk, and score points in a smooth and easy manner especially for the Center and Forward positions. Height advantage also allows you to defend your opponent and block shots that are being put up from areas closest to the net. The top-performing countries will typically not only send their most talented athletes when it comes to scoring and ball-handling but their tallest athletes as well in order to gain maximum competitive advantage. Evidence of this can be drawn from the observation in the ridgelines distribution that the median height of medalists is larger than the median height of non-medalists.*

On the other hand, we can observe from both of the plots that gymnastics has the shortest median height and the shortest athletes overall. This is because of the principle that the shorter a gymnast is, the easier it is for them to rotate in the air or spin at high speeds. It is very difficult for longer limbs and joints to handle the intensive training that comes with gymnastics. The top-performing gymnasts in the Olympics are typically the shortest in the pool. This is evident in the ridgelines distribution that clearly shows the median height of medalists being much lower than the median height of non-medalists. We can conclude from our analysis and discussion that the distribution of heights does indeed change for various sports between medalists and non-medalists.

Part 2

Question: *How has the proportion of men and women participating in the Olympics changed over time? Additionally, how has the proportion of athletes receiving medals changed over time between men and women?*

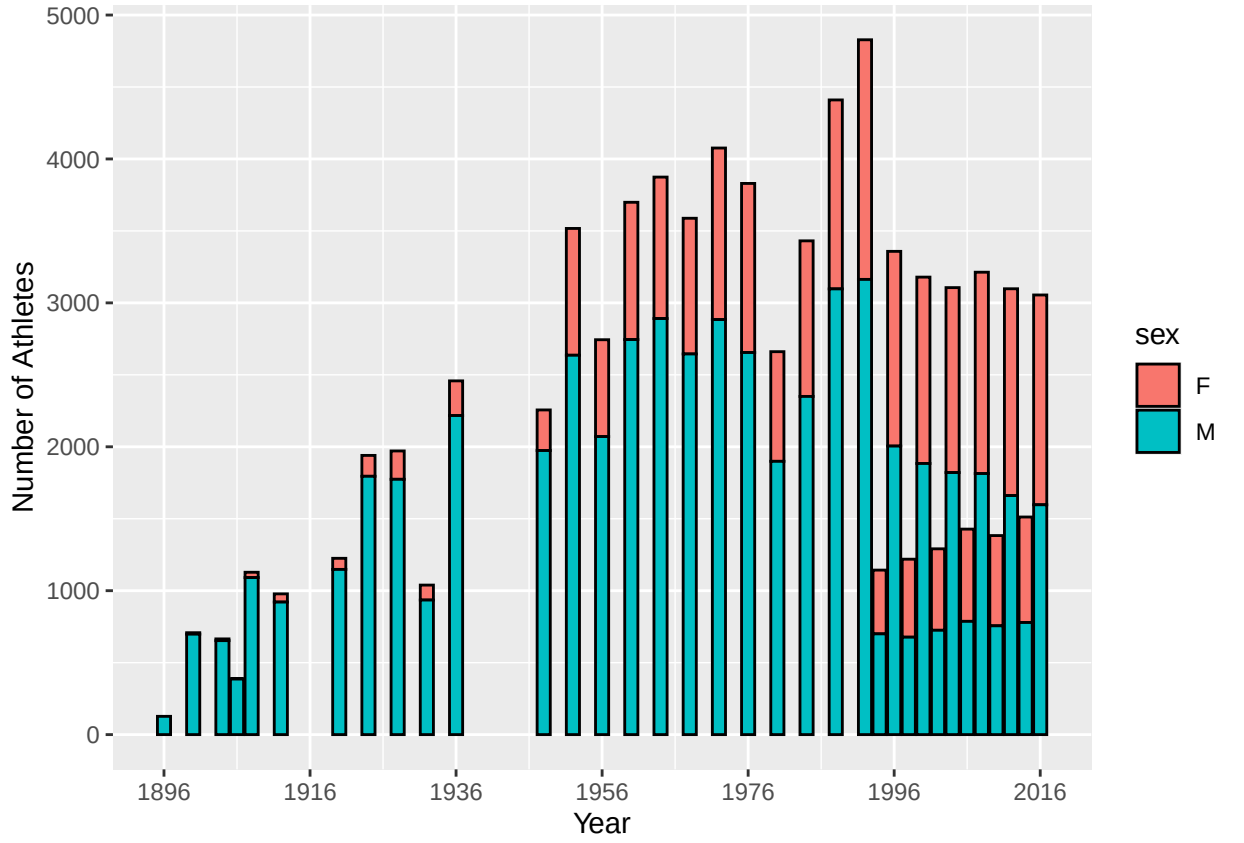
Introduction: *In this question, we are focusing on the rise in the performance and number of female athletes*

in the Olympic games since 1896. To answer this question, we will work with three variables, the athlete's `sex`*, if the athlete won a bronze, gold, silver or no* `medal`*, and the* `year`*.*

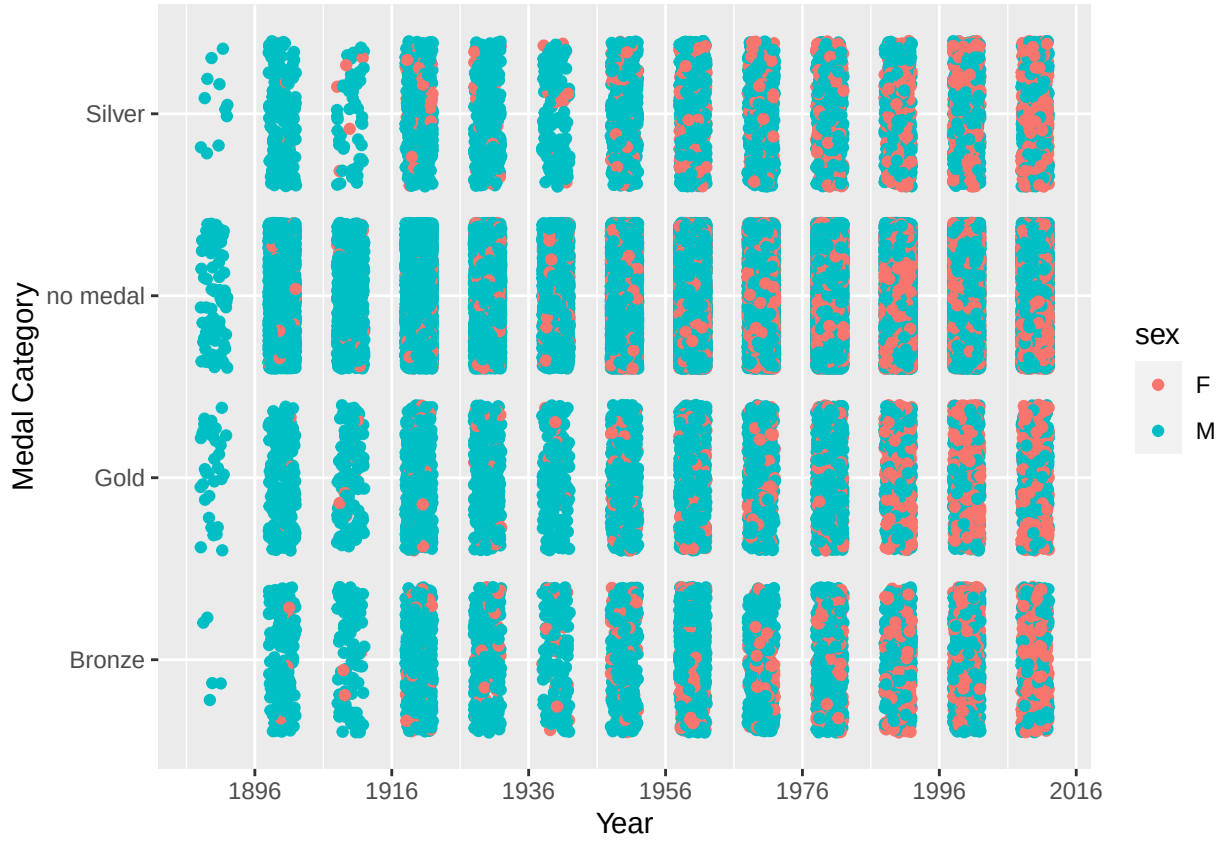
Approach: *Our approach is to use a bar graph* `geom_bar()` *to visualize the number of male and female athletes stacked together during each year of the Olympic games from 1896 up until 2016. Bar graphs are extremely useful for tracking changes over time and comparing two different groups. In order to visualize the proportion of male and female athletes receiving medals over time, we can use a jitter plot* `geom_jitter`*. Jitter plots help us to better visualize the density of the data as well as the spread of the data within groups. In our case, we will group the data points by medal status and by decade.*

Analysis: *The first plot is a bar graph that plots the number of male and female athletes over time. Note that the Winter and Summer Olympic games became separated in 1986 which is shown by the gaps in the x-axis disappearing after 1986. Additionally, the total number of athletes decreases during this time period due to the separation of the two seasons. Additionally, there were no Olympic games held during World War 1 and World War 2 which explains the two large gaps in the x-axis.*

```
ggplot(olympics_top, aes(year, fill = sex)) +
  geom_bar(color = "black", position = "stack") + scale_x_continuous(name = "Year",
             breaks = seq(1896 , 2016, 20)) + ylab("Number of Athletes")
```




```
ggplot(olympics_top, aes(x = decade, y = medal, color = sex)) +
  geom_jitter(width = 2) + scale_x_continuous(name = "Year",
      breaks = seq(1896 , 2016, 20)) + ylab("Medal Category")
```



Discussion:

Through observing the bar graph plot, we can conclude that the number of female athletes has steadily increased over time as each individual Olympic year saw a higher percentage of women competing. We can see that in 1896, the Olympic games held in Athens were exclusively for men only. Women were not allowed to compete in the Olympics until 1900. The proportion still remained heavily favorable towards men for several more Olympic games. However, events like the passing of the 19th Amendment in 1920, the massive adoption of digital technology and the Internet, and rapid increase in women-only events in the Olympics have resulted in a steady surge in the amount of female athletes. By 2016, we can see how the lengths of the red and blue bars are nearly equal. The proportion of male to female athletes goes from zero in 1896 to nearly 1:1 by the year 2016. This marks the progression of society and promotion of diversity and inclusion in the Olympic games. This also speaks volumes about the changes in attitudes towards woman over time and the increasing importance of equality and representation.

We can also observe from the jitter plot that the not only the total number but the overall performance of female athletes in the Olympic games has trended upward overtime. The proportion of athletes receiving medals between male and female athletes trended towards a 1:1 ratio over time. This is largely due to the fact that the number of Olympic events exclusively for women has drastically increased over time. Nearly 60 years ago in 1958, there were only 39 events for female athletes. By Rio 2016, the number of events for women totaled to 145, versus 161 events for men. This is evident from the largely-left-skewed jitter plot distribution which shows significant increase in red dots over the last three decades. Specifically, we can see that the density of red and blue dots appear to be slightly in favor of female athletes over the past two decades in the bronze, silver, and gold medal categories.