Project 1 - Brian Folkers (bdf676)

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:

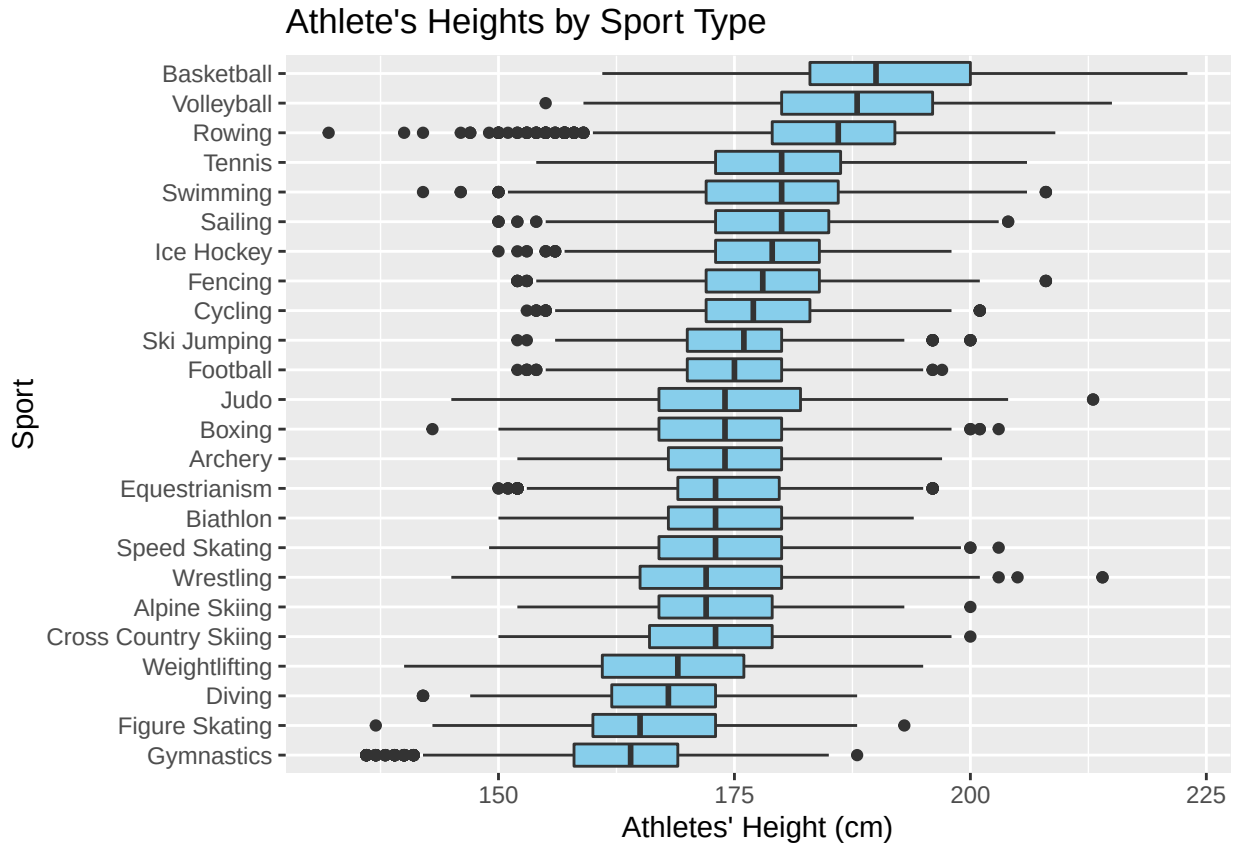
- To order boxplots by the median, you may have add the following to your ordering function to remove missing values before ordering: `na.rm = TRUE`
- To trim the tails in your ridgeline plot, you can set `rel_min_height = 0.01` inside `geom_density_ridges()`.

Introduction: This project will use the `olympics_top` dataset. This dataset is modified from the `olympics` dataset (which contains data for the Olympic Games from Athens 1896 to Rio 2016) with several added variables. The goal for this section of the project is to identify any relationship between the athletes' `height`, what `sport` they play, and their `medalist` status.

Approach: This analysis will first use a boxplot and then a ridgeline plot. The boxplot will be used to compare the distribution of athletes heights over the large number of sport types, ordered from tallest to shortest. This will allow for a basic overview of which sports have the tallest and shortest athletes. The ridgeline plot will be similar, but by separating the distributions by medalist status, it will show if the sport has a relationship between height and the likelihood of winning in that sport (the ridgeline is also faceted by `sex` to allow for a better view of the distribution of the medalists and non-medalists).

Analysis:

```
ggplot(olympics_top, aes(y = reorder(sport, height, na.rm=TRUE), x = height)) + xlab("Athletes' Height
  ylab("Sport") +
  ggtitle("Athlete's Heights by Sport Type") +
  geom_boxplot(na.rm = TRUE, fill = "skyblue")
```

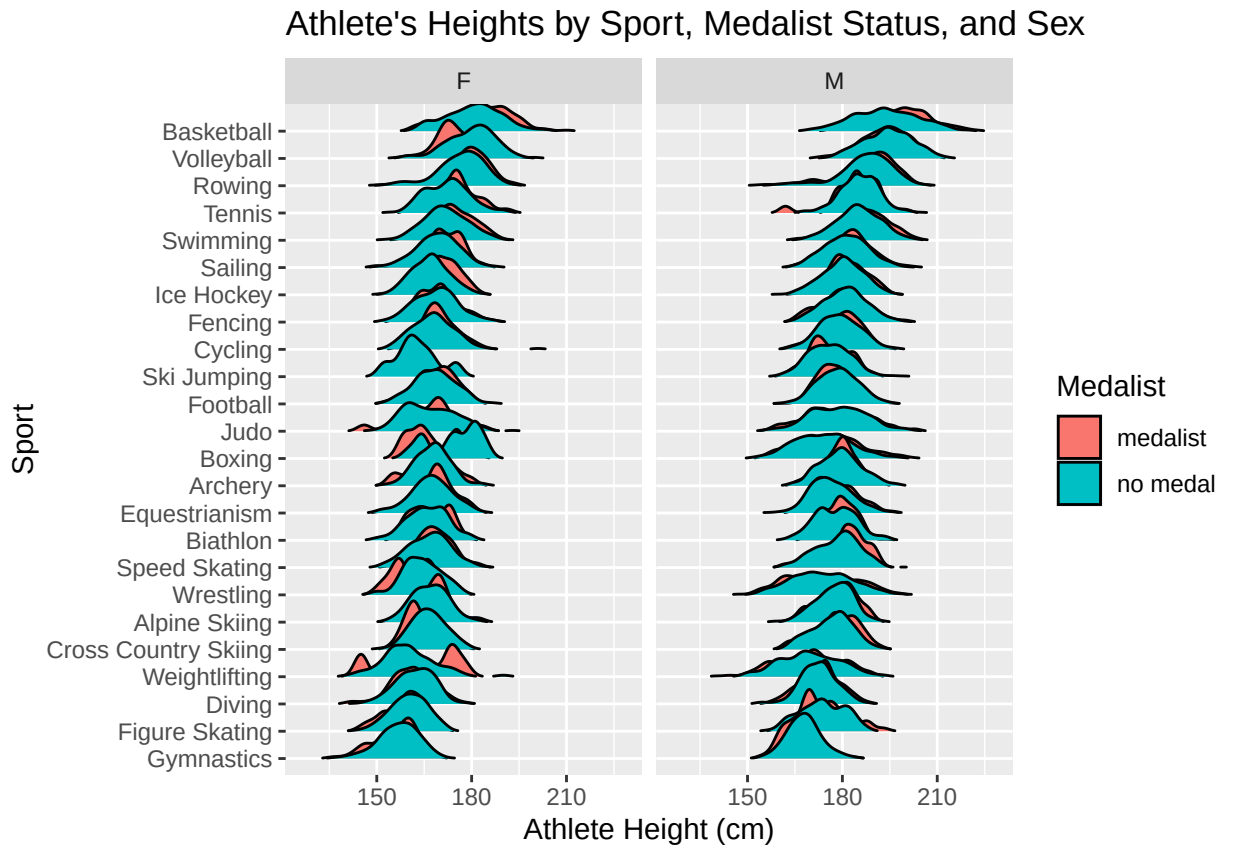



```
ggplot(olympics_top, aes(x = height, y = reorder(sport, height, na.rm=TRUE), fill = medalist)) +
  labs(fill = "Medalist") +
  xlab("Athlete Height (cm)") +
  ylab("Sport") +
  ggtitle("Athlete's Heights by Sport, Medalist Status, and Sex") +
  geom_density_ridges(rel_min_height = 0.01) +
  facet_wrap(vars(sex))
```

```
## Picking joint bandwidth of 2.16
```

```
## Picking joint bandwidth of 1.9
```

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



Discussion: The boxplot shows a clear distribution of heights for the different sport types, with basketball having the largest average height, and gymnastics having the smallest average height. These results are certianly reassonable, as a sport like basketball requires a greater height whereas gymnastics would require a lower center of gravity. The ridgeline plot shows that there is a significant difference in the heights between medalists and non-medalists for several sports. In particular, shorter females are more likely to win in wrestling, boxing, and volleyball. Interestingly, the distributions of heights for medalists and non-medalists appear to overlap considerably more for males than females. This may suggest that height is not as a significant factor in winning for males compared to females.

Part 2

Question: Does the distribution of medal types (gold, sliver, bronze) change depending on the athletes' country or age?

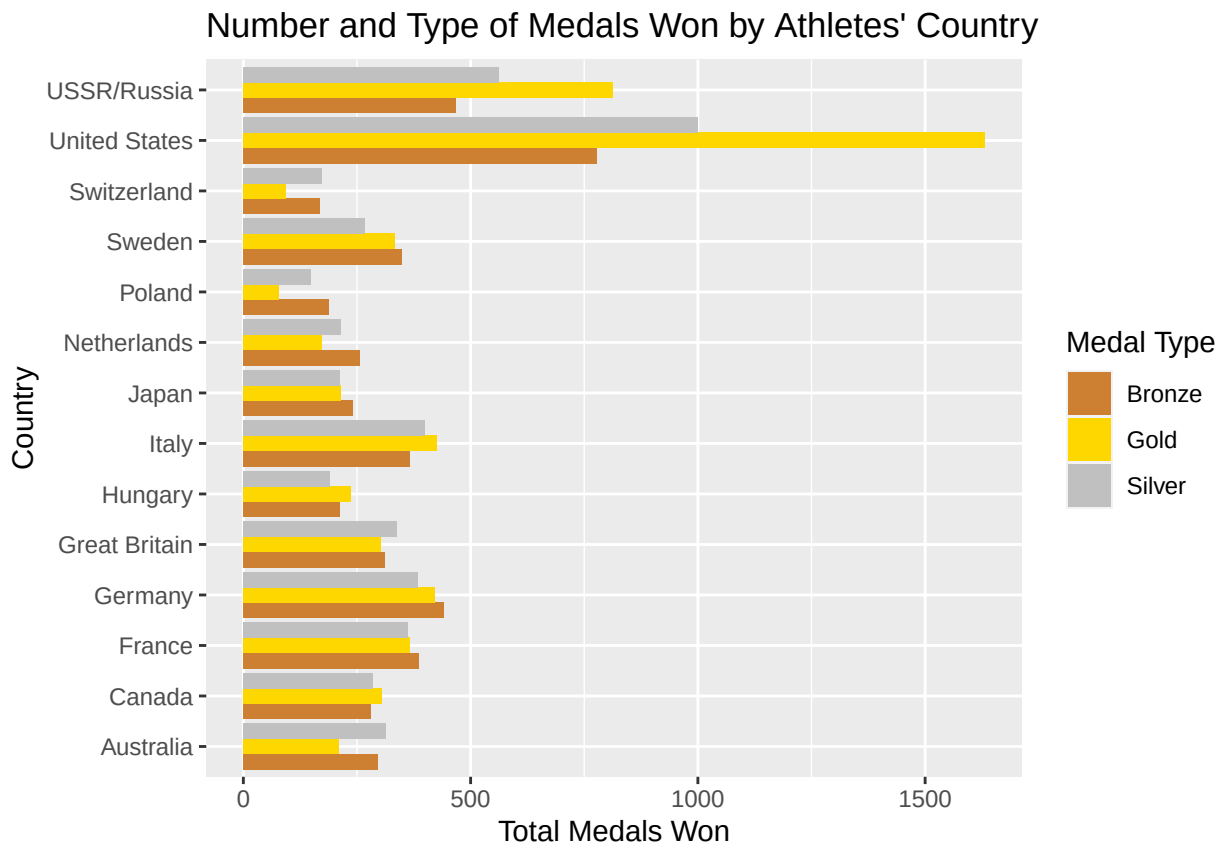
Introduction: This section of the project will used a modified form of the `olympics_top` dataset called `olympics_top_2`. This dataset removes the "no medal" category from the `medal` variable so as to view the data in terms of gold, silver, and bronze only. The goal for this section of the project is to identify any relationship between the athletes' country they play for (`team`), what `age` they are, and their `medal` they won.

Approach: This analysis will first use a grouped bar chart and then a histogram. The bar chart will be used to show the distribution of the different types of medals won for each country in the dataset. This will allow for the visualization of the amount the different types of medals for each country. The histogram will show the age of all athletes while being faceted by type of medal won. This will allow for the visualization of the age distribution for each medal type.

Analysis:

```r
olympics_top_2 <- olympics_top %>% filter(medal != "no medal")

ggplot(data=olympics_top_2, aes(y= team, fill = medal)) +
  labs(fill = "Medal Type") +
  xlab("Total Medals Won") +
  ylab("Country") +
  ggtitle("Number and Type of Medals Won by Athletes' Country") +
  geom_bar(position = "dodge") +
  scale_fill_manual(values = c("#CD7F32", "#FFD700", "#C0C0C0"))
```




```r
ggplot(olympics_top_2, aes(age)) +
  geom_histogram(fill = "skyblue", binwidth = 3) +
  xlab("Athletes' Age (years)") +
  ylab("Total Medals Won") +
  ggtitle("Number and Type of Medals Won by Athletes' Age") +
  facet_wrap(vars(medal))
```

```
## Warning: Removed 109 rows containing non-finite values (stat_bin).
```

Discussion: The bar chart shows that medal type distribution generally does vary by country. The United States and the USSR/Russian Federation are unique for having considerably more Gold medals than any other type compared to the other countries. Most other countries also have a proportional number of medal types, often with few Gold medals. This might suggest that most events are eventually won by either the US or Russia. For the histogram, it would appear that age does not vary considerably between type of medal won. There does appear to be a larger number of gold medalists around their mid-twenties, but the distribution is similar for all medal types. This might suggest that age does not significantly effect the type of medal won.