# Project 1

**Jason Flores UT EID:** *jf36995*

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:**

We are working with the `olympics` dataset which contains 134,731 athletes that have competed in the Olympics since Athens 1896 to Rio 2016. In this data set `olympics`, each row contains the name of the athlete that have competed with 13 variables that explain their sex, age,height (cm), weight (kg), team they are representing, the national Olympic committee region, the year of participation, the type of olympics game name, the season, the host city, the sport, the event, and what medal they won. Additionally, we will be working with the dataset `olympics_top` which is derived from `olympics` dataset but contains 4 new variables, which are the decade the olympics took place, if they won gold, if they were a medalist, and what type of medal they won

To answer the 1st part of question 1, we use data set `olympics` and the variables we will be working on is the sport (column `sport`), and the height (column `height`). No additional variables are needed since we only care about finding out which sport has the tallest/shortest athlete To answer the 2nd part of question 2, we use data set `olympics_top` and the variables we will be focusing on are sport (column `sport`), height (column `height`), and whether they won a medal or not (column `medalist`)

**Approach:**

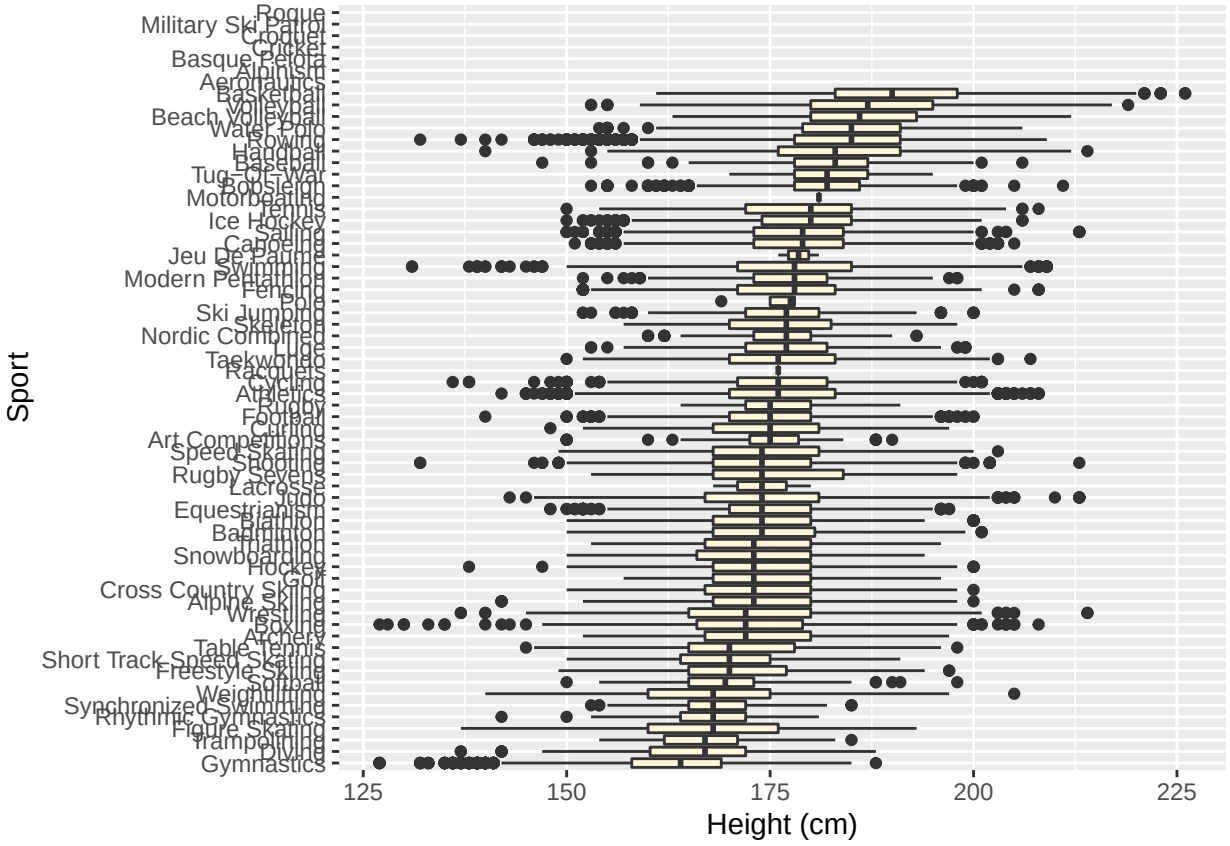
My approach in answering the 1st part of question 1, is using a `boxplot` and organize the height using `fct_reorder`. This will show a clear visualization on what sport had the lowest/highest median height. The only issue is the graph may be clustered.

Furthermore, the approach in answering the 2nd part of question 1 is to use a `ridgeline plot` to compare the `height` between medalist and non-medalist using `geom_density_ridges`.

**Analysis:**

```
#Finding the number of athletes that competed from 1896-2016 using *table(olympics$name)*
#To answer the 1st part of the question we working with Olympics data set
#Using bloxplots to determine which sport has the tallest/shortest athlete
# Sports on y-axis, height (cm) on x-axis
ggplot(olympics,
       aes(height,fct_reorder(sport,height, na.rm = TRUE))) +
  geom_boxplot(fill = "cornsilk") +
  scale_x_continuous(
    name = "Height (cm)")+
  scale_y_discrete(
    name = "Sport")
```

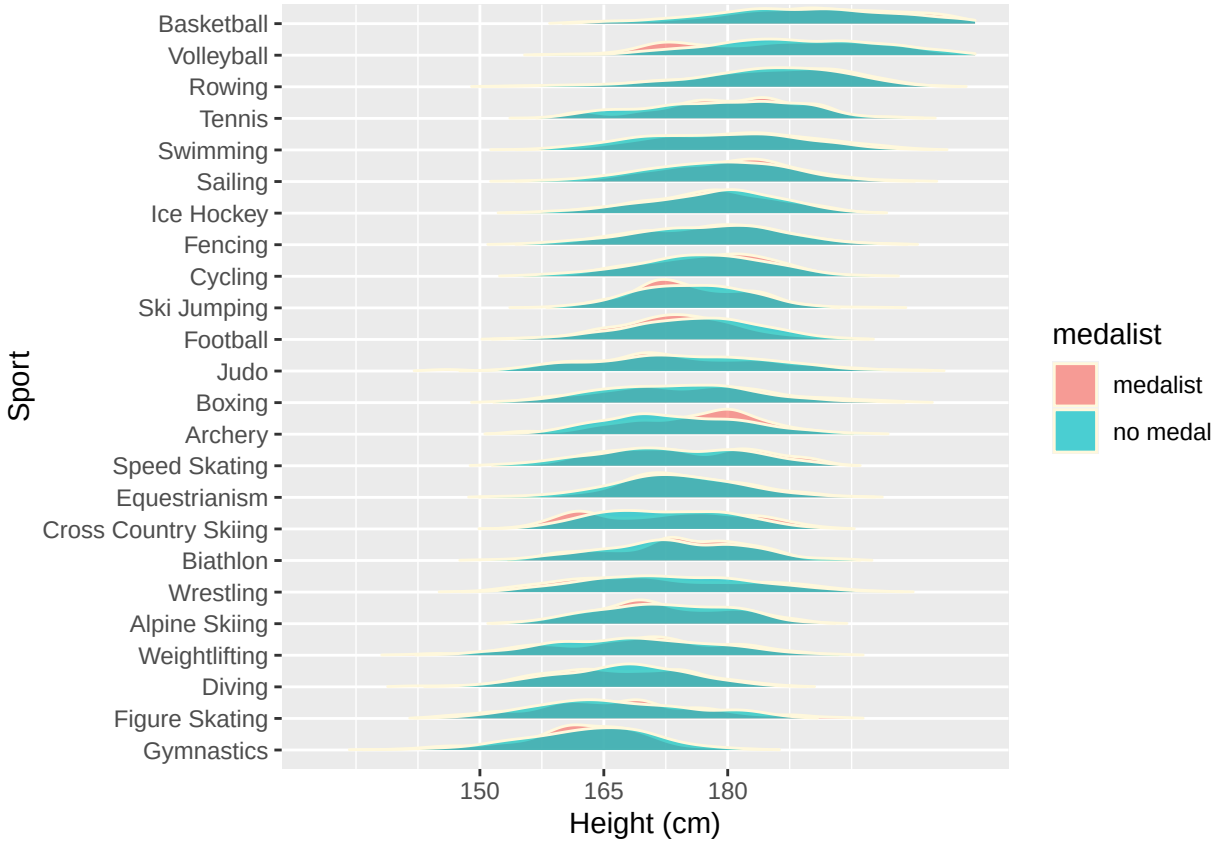
```
## Warning: Removed 60171 rows containing non-finite values (stat_boxplot).
```




```
#using a ridgeline plot
ggplot(olympics_top,
       aes(height,fct_reorder(sport,height, na.rm = TRUE), fill = medalist)) +
  geom_density_ridges(rel_min_height = 0.01, color = "cornsilk", alpha=0.7, scale = 0.9) +
  scale_x_continuous(
    name = "Height (cm)",
    limits = c(130,210),
    breaks = c(150, 165, 180),
    labels = c("150", "165", "180"))+
  scale_y_discrete(
    name = "Sport")
```

```
## Picking joint bandwidth of 2.18
```

```
## Warning: Removed 19190 rows containing non-finite values (stat_density_ridges).
```




**Discussion:**

To answer the 1st part of question 1. The sport with the shortest athletes is Gymnastics and the sport with the highest athletes is Basketball. A bit weird that other sports appeared basketball, not sure what that is about.

To answer the 2nd part of question 1. No, overall, the height distribution between medalist and non-medalist for each sport appear to be the same.

## Part 2

**Question:**

Does age distribution vary on the type of medal won? And which team has the most gold-medalist?

**Introduction:** We are working with the dataset `olympics_top`. As mentioned in ***Part 1*** the dataset is derived from `olympics` but it has 4 new variables being `medalist` (medalist/non-medalist), `medal` (gold,silver,bronze, no medal), `gold` (no gold/gold medalist), and `decade` (the decade the Olympics took place).

To answer the 1st part of question 2, the variable we will be focusing on in the dataset `olympics_top` is the age (column `age`) and medal (column `medal`). We do not need any other variables since we care about the median age.

To answer the 2nd part of question 2, the variable we will be focusing is team (column `team`) and gold (column `gold`). No additional variables needed as we only care about gold and the team.

**Approach:**

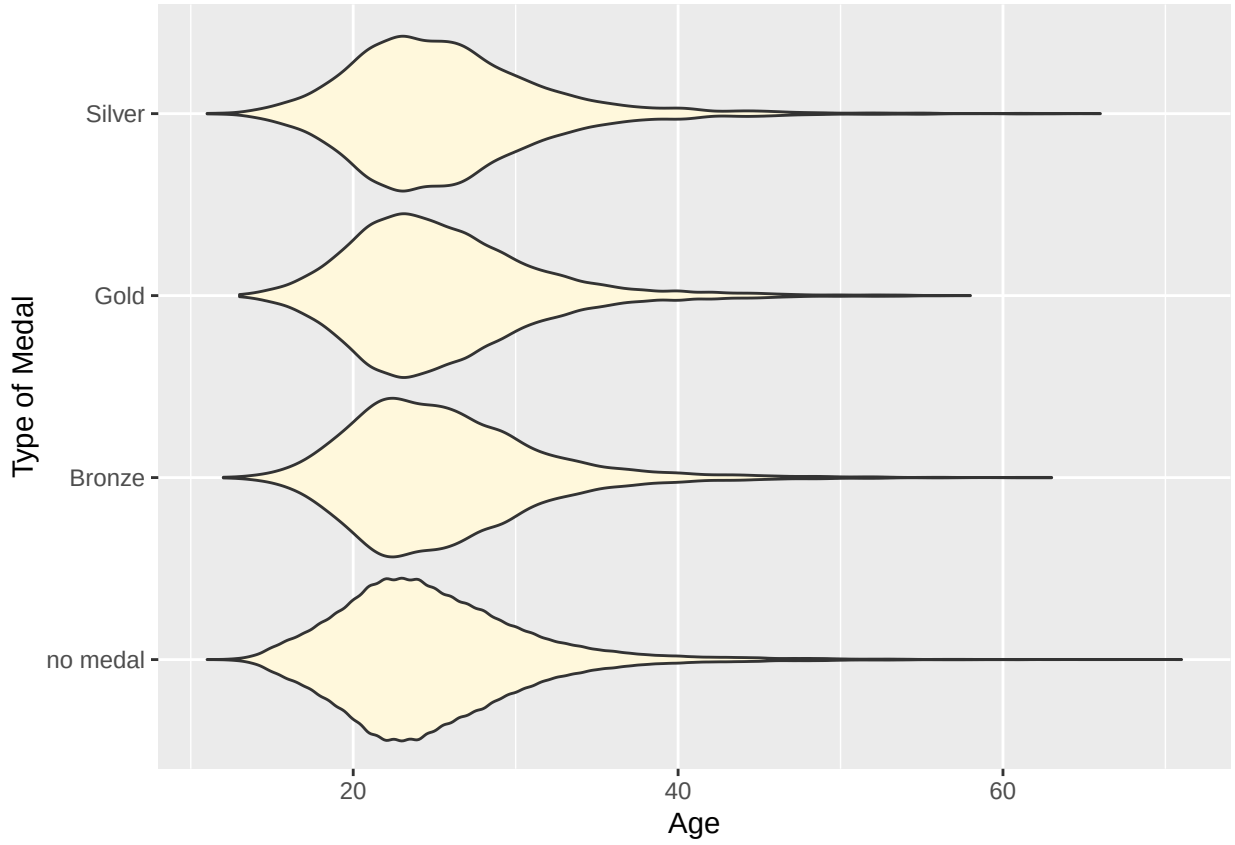
The approach to answering the 1st part of question 2 is to use a `violin` plot `geom_violin` and organize the age using `fct_reorder`. This method is similar to a `boxplot` but using a different visualization to show the data. In the graph, we will be able to see any difference if there's any for age across the medals won.

The approach to answering the 2nd part of question is to use a `barplot` (`geom_bar`) we will be using `facet_wrap` to view the teams that won the most gold medals.

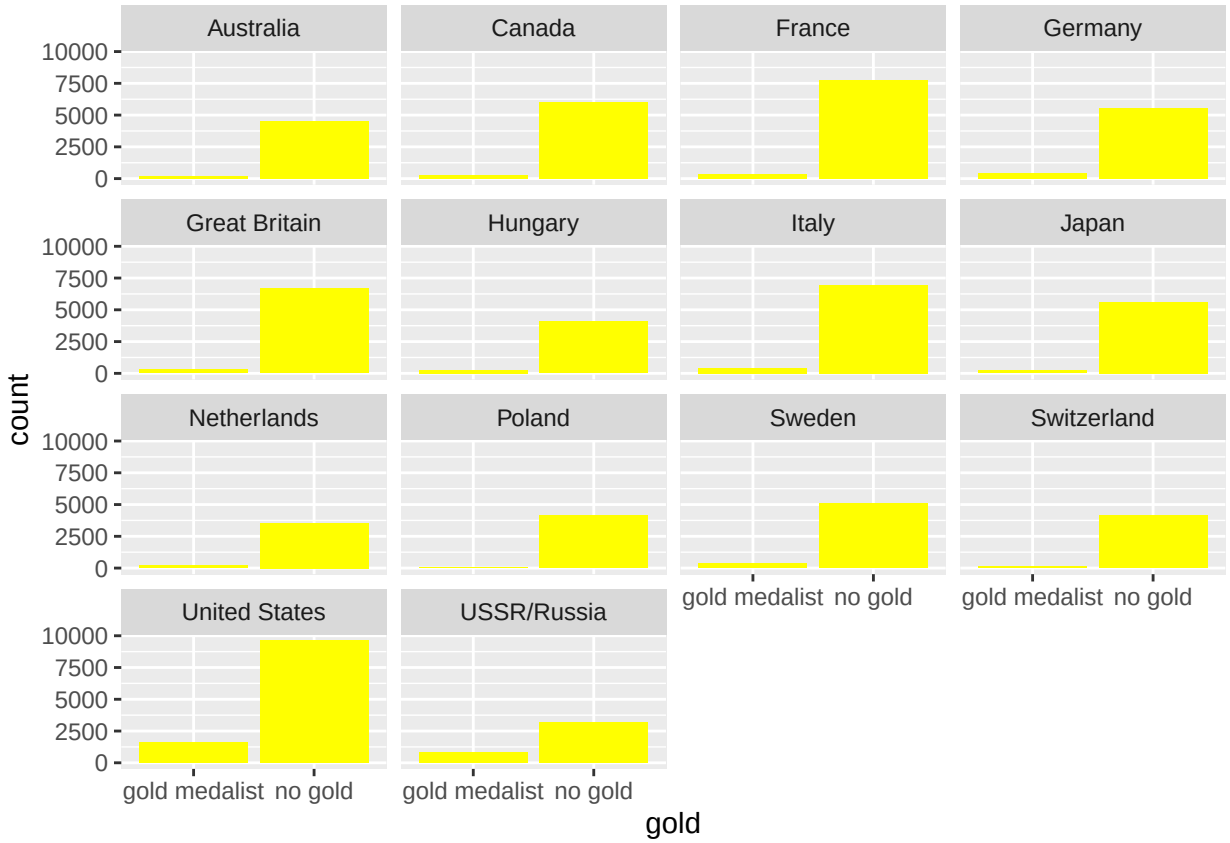
**Analysis:**

```
# Using a violin plot
ggplot(olympics_top,
       aes(age,fct_reorder(medal,age, na.rm = TRUE))) +
  geom_violin(fill = "cornsilk") +
  scale_x_continuous(
    name = "Age")+
  scale_y_discrete(
    name = "Type of Medal")
```

```
## Warning: Removed 2421 rows containing non-finite values (stat_ydensity).
```




```
# Using a barplot
ggplot(olympics_top, aes(gold)) +
  geom_bar(fill = "yellow") +
  facet_wrap(~team)
```




**Discussion:**

The answer to 1st part of question 2 is that there seems to be no difference in the age distribution on the type of medal won. It looks like an equal distribution across the mean age and medals won.

The answer to 2nd part of question 2 is United states is the team with the most gold-medalist.