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()`.

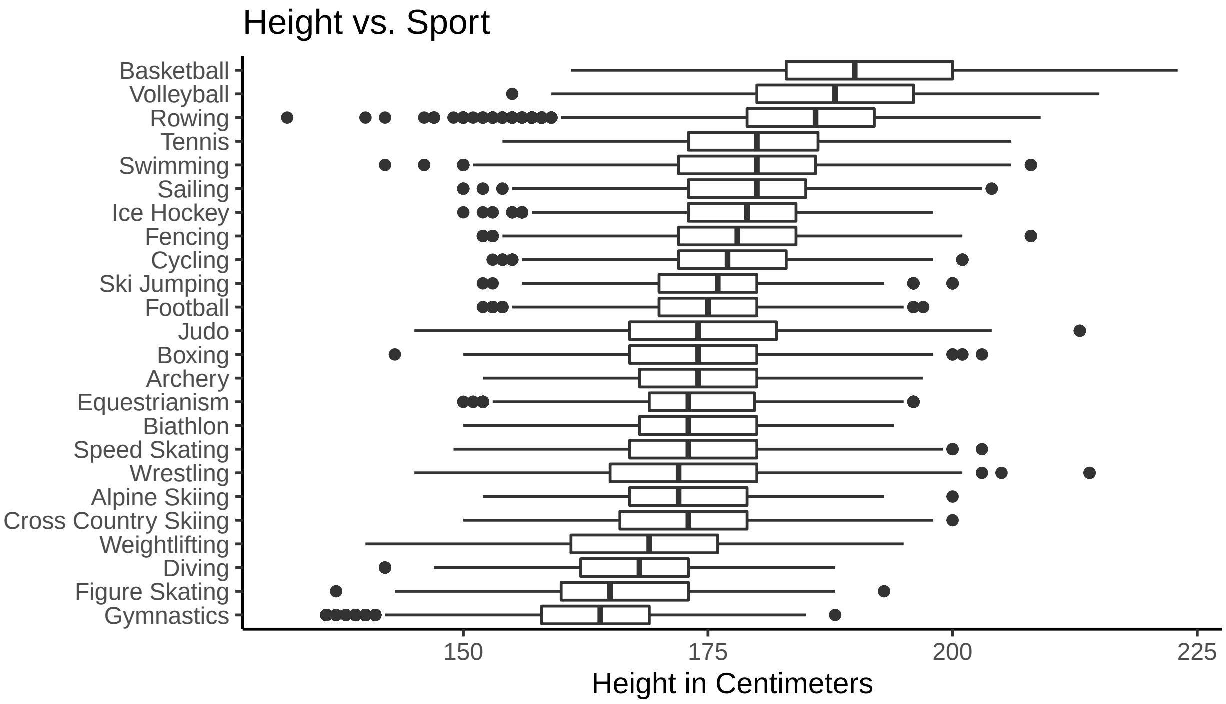
Introduction: *We are charged with determining the tallest or shortest Olympic athletes from the 1896 Athens games to the 2016 Rio games, using the data set "Olympics_top".To determine which Olympic sports have the tallest or shortest athletes, first we must consider the variables "sport" and "height". The variable "sport" is a categorical nominal variable and the variable "height" is a numerical ratio variable.In order to determine the changes of height distributions between medalists and non-medalists, we need to consider the variables "height", "sport", and the categorical variable "medal".*

Approach: *The variable "sport is nominal variable and the variable height is a continuous numerical variable. This combination of variables creates an environment that is suitable for a box plot as it is able to show many distributions events side by side, thus creating an easy comparison environment. In order to determine the distribution of heights based on event and height, we must use a ridge line plot. This will allow us to see and compare multiple distributions at the same time easily.*

Analysis:

```r
# This code creates a box plot which compares the Olympic athletes height's by sport.
df<-olympics_top
ggplot(df,aes(y=reorder(sport,height,na.rm=T),x=height))+
  geom_boxplot()+
  theme_classic()+
  xlab("Height in Centimeters")+
  ylab(NULL)+
  ggtitle("Height vs. Sport")
```

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



```
# This code creates a faceted ridge line plot, which compares Olympic athletes heights by sport and whe
ggplot(df,aes(y=reorder(sport,height,na.rm=T),
              x=height,fill=medal))+
  geom_density_ridges(rel_min_height = 0.01,alpha=0.7)+
  theme_linedraw()+
  facet_wrap(vars(medal))+
  xlab("Height in Centimeters")+
  ylab(NULL)+
  ggtitle("Height Distribution in Olympic Sports")
```

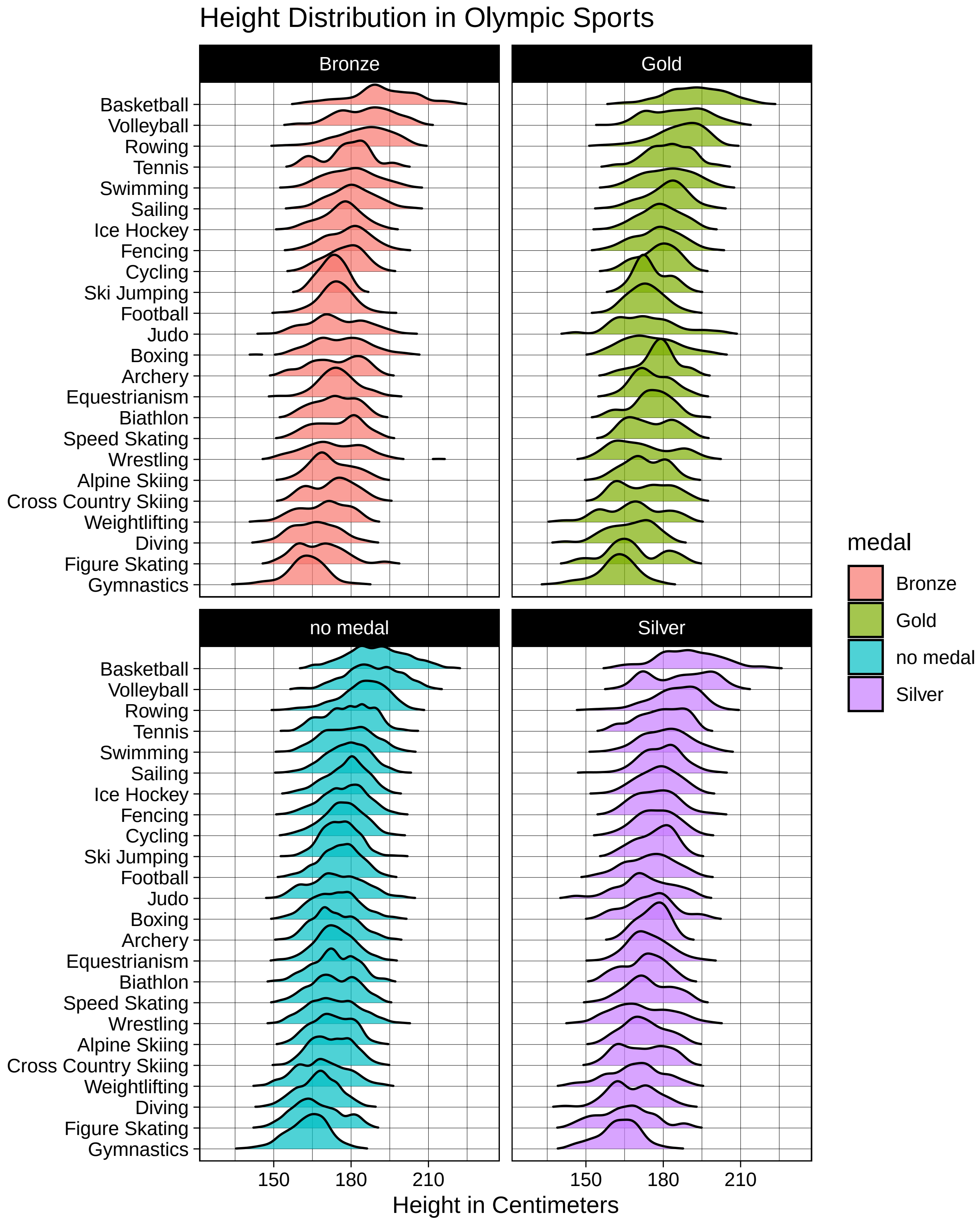
```
## Picking joint bandwidth of 3.05
```

```
## Picking joint bandwidth of 3.12
```

```
## Picking joint bandwidth of 1.85
```

```
## Picking joint bandwidth of 3.12
```

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



Discussion: *From the box plot a viewer can easily see that the sport with the greatest median height is basketball and the sport with the least median height is gymnastics. This means we expect to see on average the tallest Olympic athletes playing basketball and the shortest Olympic athletes competing in gymnastics.In the ridge line plot viewers can see that the height distributions for between the medalists and non-medalists in the Olympic sports is not uniform. Thus, the medal winning Olympic athletes have different heights than the* *non-medalists.*

Part 2

Question: *Which sports have athletes of the greatest mean age and the smallest mean age? Is the age distribution for the Olympic sports different between men and women?*

Introduction: *To determine which sports have athletes of the greatest age and smallest age, we must consider the numerical ratio variable "age" and categorical nominal variable "sport". In order to determine if there is a difference in age distribution in Olympic sports between men and women, we must consider the categorical nominal variable "sex", and "age"*

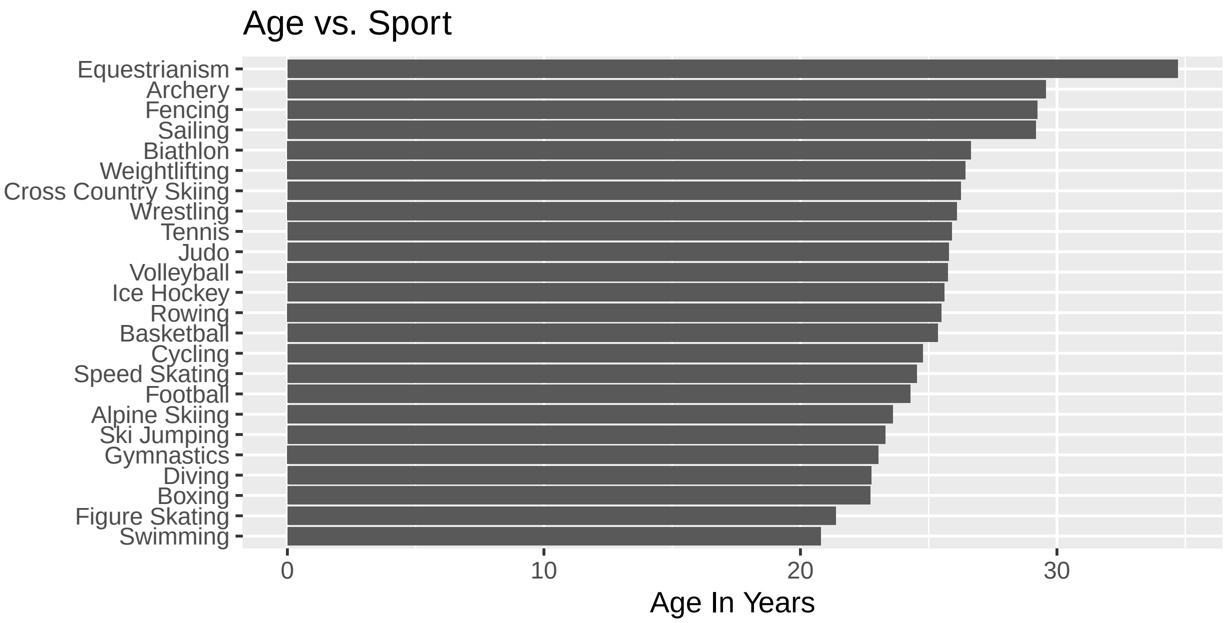
Approach: *In order to determine which sports have athletes of the greatest and least mean age, we will use a histogram. A histogram will allow us to map the categorical variable "sport" on the y-axis and the numerical variable "age" on the x-axis. This configuration once arranged by mean age, will allow a viewer to quickly and easily determine the greatest and least mean sport ages. To determine whether men and women have different age distributions in different sports, we must map the variable "sex"and "age" on an overlapping density plot.This will allow the viewer a more in depth view to see the differences of age by sex in the many Olympic sports.*

Analysis:

```r
# This code creates a histogram comparing the mean ages of Olympic athletes in different sports
ggplot(df,aes(x=age,y=reorder(sport,age,na.rm=T)))+
  geom_histogram(stat = "summary",fun="mean")+
  xlab("Age In Years")+
  ylab(NULL)+
  ggtitle("Age vs. Sport")
```

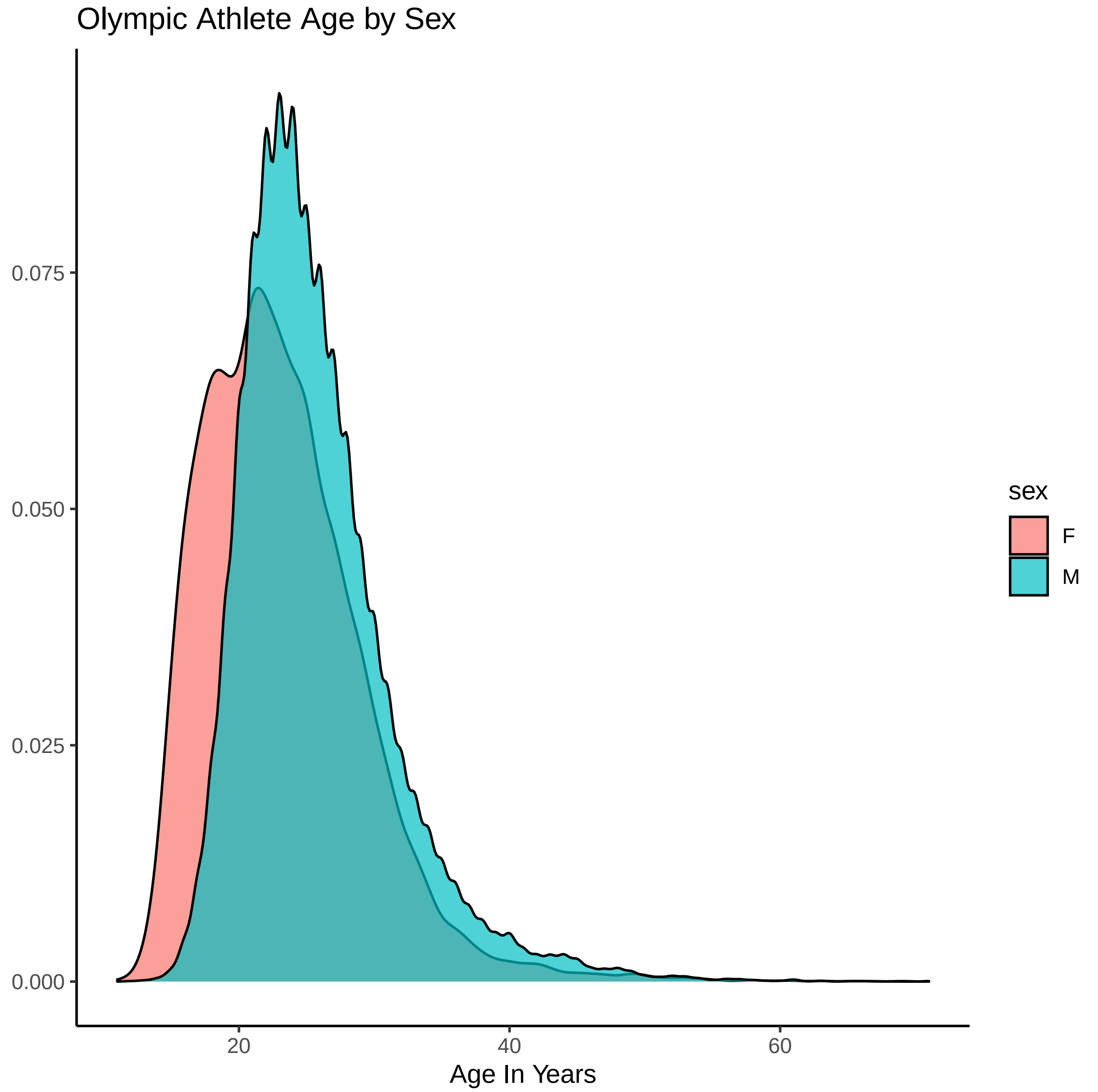
```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```

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



```r
# This code creates many plots showing the age distribution of men and women in Olympic Sports
ggplot(df,aes(x=age,fill=sex))+
  geom_density(alpha=0.7)+
  xlab("Age In Years")+
  ylab(NULL)+
  ggtitle("Olympic Athlete Age by Sex")+
  theme_classic()
```

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



Discussion: *In the first plot a viewer is able to quickly discern that the Olympic sport with the greatest mean age is Equestrianism, and the sport with the least mean age is swimming.In the overlapping density plot, a viewer can see that the age distribution for men and women in the Olympics was approximately the same, but men yielded a slightly greater mean age.*