# 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:** *In this project, the author was provided with an olympic games dataset entitled "olympics_top" from a webpage with the following URL: " 'https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/data/2021/2021-07-27/olympics.csv' ". This dataset contains information on each athlete that has competed in the winter and/or summer olympic games from the Athens competition in 1896 to the Rio competition in 2016. The dataset contains records on qualitative data such as the athlete name, sex, team, nationality, specific olympic competition, olympic season, host city, sport, event name, medal award for each of the athletes. The dataset also contains records on quantitative data including the age, height, and weight of the athletes.*

*In Part 1, the author was first asked to examine the relationship between the height of the olympic athletes and their associated sport. To accomplish this, the qualitative data specifying the sports, and quantitative data specifying the athlete's height (in centimeters) was needed. The second objective of Part 1 was to examine how the distribution of athlete heights changed from medalists to non-medalists according to the sport.*

**Approach:** *For the first question in Part 1, the author plans on using boxplots to visualize the relationship between athlete height and participating sport. The reasoning for using boxplots here is that they are great for visualizing the distribution of quantitative data. In particular, they depict a line characterizing the median of the quantiative data which is very useful when the purpose of visualizing data is to determine how the mean or median value (height) changes from one group (sport) to another. The author will use the following code framework to generate this visualization: "ggplot(dataframe, aes(x=reorder(y,x, na.rm = TRUE), height)) + geom_boxplot(fill ="color") + coord_flip() + xlab("") + ylab("").*

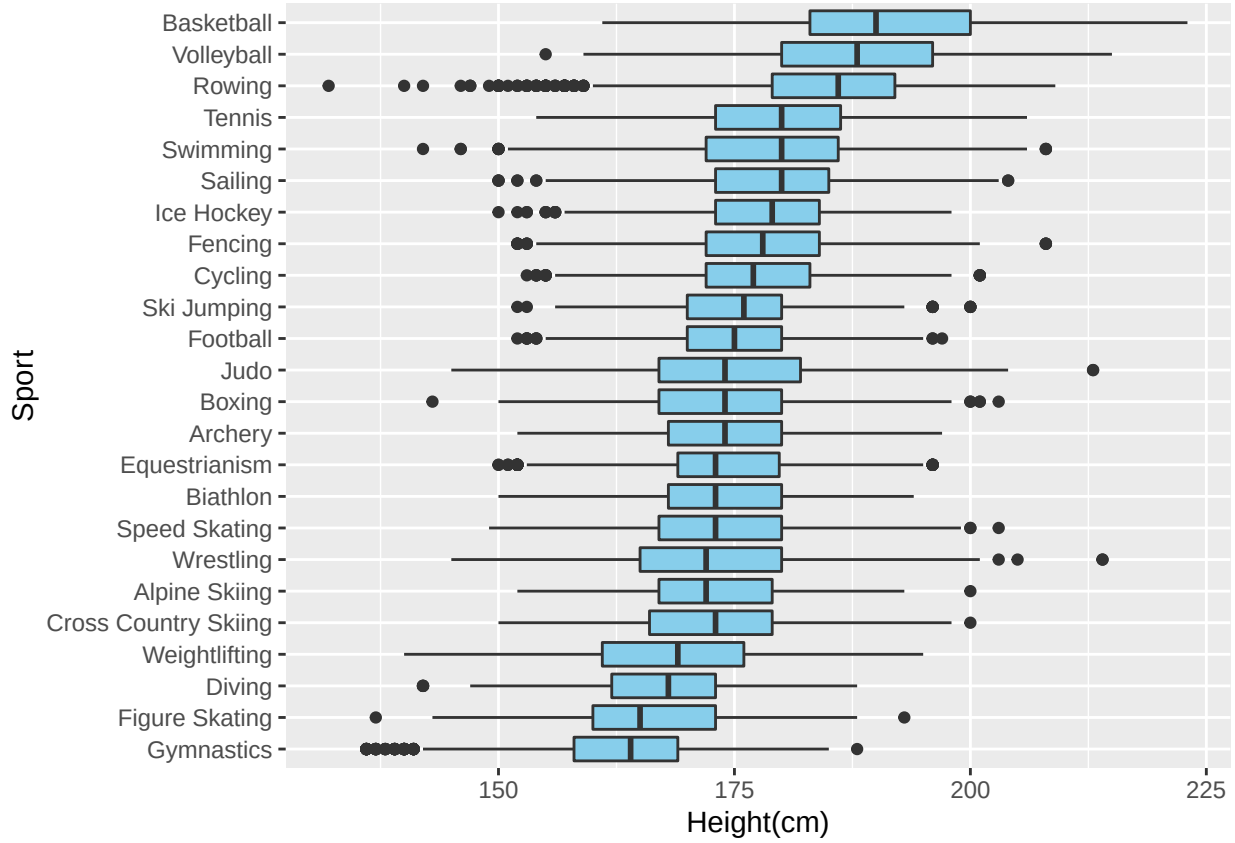
*For the second question in Part 1, the author plans on using ridgeline density plots to visualize the change in the distribution of heights for each sport according to athletes that medal, and those who do not. As an aside, boxplots are limited to single modality of their distributions, namely, it is not possible to depict multimodal distributions through datasets. Conversely, ridgeline density plots cluster frequencies of quantitative data into sub-groups(ranges of quantitative data) for each category like histograms and use interpolation to generate smooth curves that connect the data points for each sub-group. The height of these smooth curves*

*is proportional to the count/frequency of the quantitative variable in each sub-group. This allows for the user to more easily visualize multimodal distributions. Another advantage of using ridgeline density plots in "ggplot2" is that multiple density plots can be generated for each category of a qualitative variable (discrete aesthetic) allowing for the user to visualize how the distribution of data varies across multiple groups within each overlying qualitative category. Since we are interested in visualizing how the distribution of a quantitative variable (height) varies for each qualitative category (sport) according to multiple groups (medalists, and non-medalists), it is reasonable to use ridgeline density plots.*

**Analysis:**

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

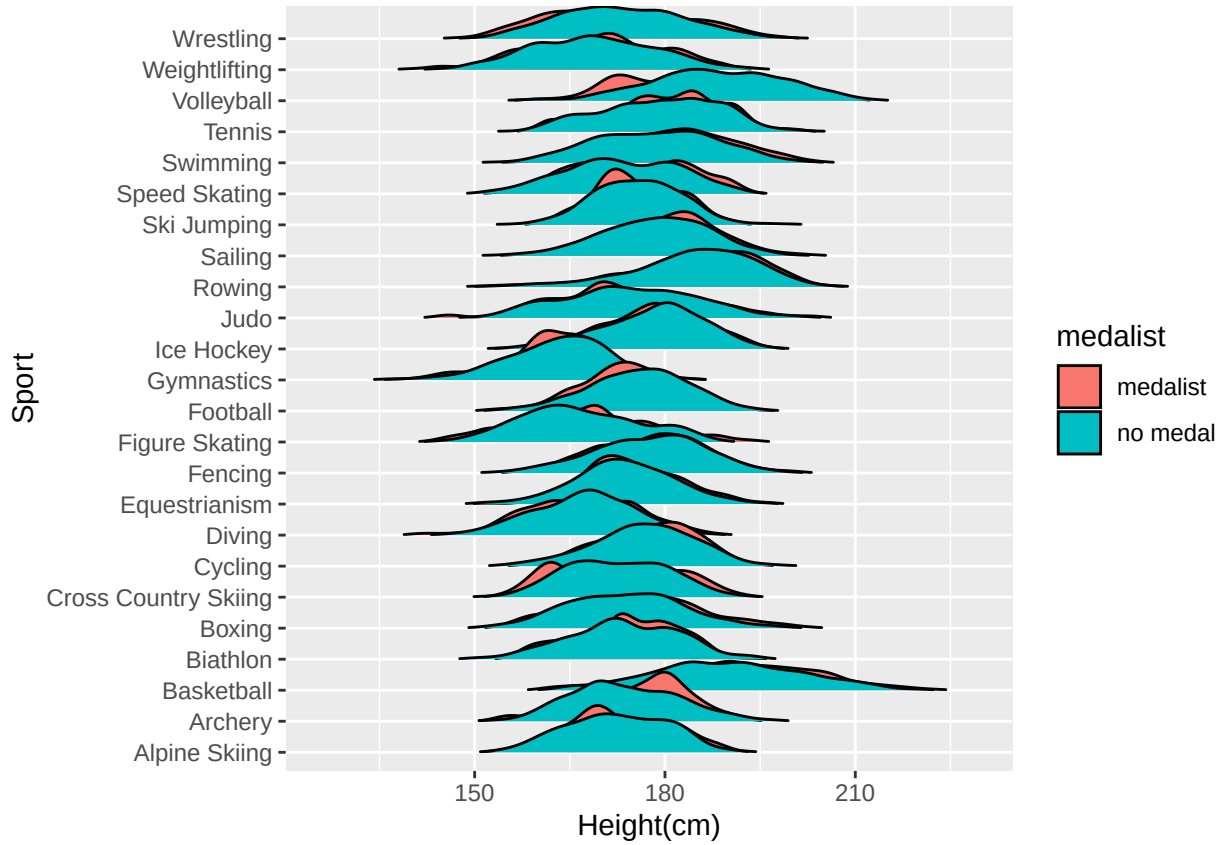
```
## Warning: Removed 19103 rows containing non-finite values (stat_boxplot).
```




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

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

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




**Discussion:** *With reference to the boxplot visualizations used to depict the distributions of athlete height according to their participating sport, it is clear that gymnastics athletes have the lowest median height while basketball athletes have the highest median height. Additionally, the spread of the height data for the gymnasts is one of the lowest across the sports whereas the spread of the height data for the basketball athletes appears to be the highest. To understand why this is the case, let us first digest attributes that many successful gymnasts and basketball athletes have in common.*

*With respect to gymnastics, these athletes are scored on their ability to perform elegant manuevers with their upper and lower body through twists, turns, spins, and other movements that necessitate a strong core, an athletic body, and most importantly, high flexibility. While primarily being associated with muscle tone, flexibility is also related to height. Generally, the shorter the height an athlete has, the greater their flexibility. It is also important to note that height is proportional to body weight. From classical mechanics, a body of a large height and high body weight will have a much greater moment of inertia (greater resistance to control change in motion) than one of a smaller height and lesser body weight making it more difficult for the former to perform complicated manuevers that require body control and changes in motion in gymnastics. Thus, it is reasonable to conclude that gymnastics athletes have a very small median height because height is inversely proportional to flexibility and directly proportional to body weight. Additionally, it is rational that the plot reflects the spread of the gymnastics height data as amongst the lowest of the sports because if gymnastics athlete is too short, then he/she won't be able to generate enought power to achieve high lift for the manuevers, and if he/she is too tall, then the aforementioned body weight and flexibility concerns become apparent.*

*Looking at the ridgeline plots, we see that the plot depicting the distribution of gymnast heights for medalists peaks at a lower height value than that of the plot that depicts the distribution of gymnast heights for non-medalists. This is reasonable as the top-performing olympic teams tend to send their shortest, and most-flexible gymnastic athletes to compete in the biggest stages of the sporting competitions.*

*With reference to the boxplot that visualizes the distribution of basketball heights, it is common knowledge that height differentials amongst basketball players facilitate many advantages on the court. Generally, the*

*more height a basketball player has, the easier it is for the athlete to defend (through rebounding and blocking shots), and score (as the taller an athlete is, the closer he/she is to the basket). From this information, it is reasonable that the median height of basketball athletes is very high according to the boxplot distribution as height is directly associated with defensive and scoring abilities. Now, with that being said, the tallest basketball athletes tend to play the positions that are closest to the basket (centers and power forwards), where height differential is the most important for being able to grab rebounds and score in close proximity to the basket, the shortest basketball athletes (point guards, shooting guards, and small forwards) tend to be the quickest and most athletic and are able to impact the game through scoring and passing further away from the basket. From this, it follows that the height differential between the two groups tends to be quite high. Although this height differential is typically very high, the median height of the shorter group of basketball athletes is still moderately above the median height of adults in general, and the median height of the taller group is well above the median height of the general adult population. From this, we can conclude that the high median, and large spread of the basketball player heights as reflected in the boxplot is reasonable.*

*Again, looking at the ridgeline plots, we see that the plot depicting the distribution of basketball athlete heights for medalists peaks at a slightly higher value than that of the non-medalists. This is sensible because the highest-performing international basketball teams tend to be comprised of the tallest athletes who possess major advantages in key performing areas including rebounding, shot blocking, and scoring around the rim.*

*In conclusion, it was determined that the sport of basketball, and gymnastics have the tallest and shortest athletes respectively according to the boxplot distributions provided in the first figure. Using the ridgeline plots, it was also determined that there is a noticeable median height differential between gymnastic athletes that do and do not medal, and between basketball athletes that do and do not medal. In particular, gymnastic athletes that medal tend to be shorter than their counterparts who do not medal, and conversely, basketball athletes that medal tend to taller than those who do not. Lastly, we can conclude that the distribution of heights for all sports generally changes between medalists and non-medalists.*

**Part 2**

**Question:** *Which sports have the youngest or oldest athletes? And how does the distribution of age change for female medalists, female non-medalists, male medalists, and male non-medalists? .*

**Introduction:** *In Part 2 of the project, the author seeks to use trends presented in data visualizations to determine which sports (if any) have, on average, the oldest and youngest participating athletes. To accomplish this, the data in the "age" and "sport" columns must be extracted from the "olympics_top" data frame. In a related study, the distribution of athlete age for categories that lie at the intersection of medalists, and sex will be examined. This will be achieved using data from the "age", sex" and "medalist" columns in the "olympics_top" data frame.*

**Approach:** *To investigate the first part of the question, the author will use violin plots to visualize the distribution of athlete age for each sport. Violin plots are often used as an alternative to boxplots to display the distribution of quantitative data. Although it is not possible to identify certain measures of the spread (such as interquartile range), the center (median), or precise outliers, the shape of violin plots reflect the qualitative nature of the spread and center of the distribution. In particular, one can determine how the frequency of the data changes with change in the value of the quantitative variable by examining the thickness of the plot. To expand, the thickness is proportional to the frequency. An advantage of using violin plots is that it is possible to visualize multimodal data (as with ridgeline plots), a feature that is not available in boxplots. Since the data may have multiple modes, and since we are interested in distinguishing between the sports that have the youngest athletes from those that have the oldest at a high-level, it is reasonable to use violin plots.*

*For the second part of the question, the author will generate histogram subplots using the facet wrap feature to compare the distribution of age across the aforementioned four categories. The facet wrap feature is desirable as it allows the author to visualize the data in distinct plots side-by-side on the same figure. This eliminates overlap and clutter of data which could arise from using other forms of visualization where all data is presented on the same plot. It follows that using histogram subplots to depict the distribution of age in each of the four categories is the best approach to understand how age frequencies vary across the subgroups because*

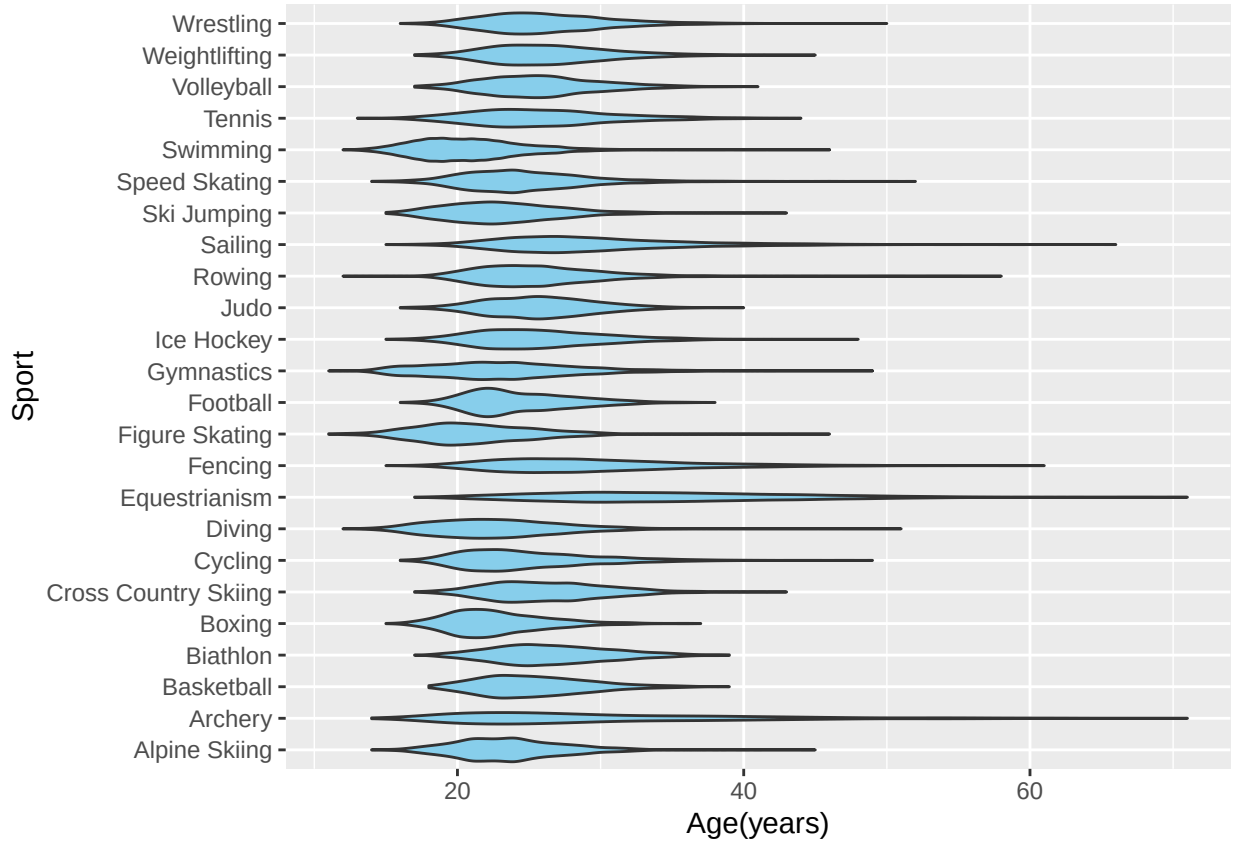
*histograms are an efficient way to visualize multimodal quantitative data and because facet wraps allow for multiple distributions to be compared side-by-side with no clutter or overlap.*

**Analysis:**

```
# Generation of Violin Plot to depict distribution of athlete age according sport

ggplot(olympics_top, aes(x = age, y = sport)) +
  geom_violin(fill = "skyblue") + xlab("Age(years)") + ylab("Sport")
```

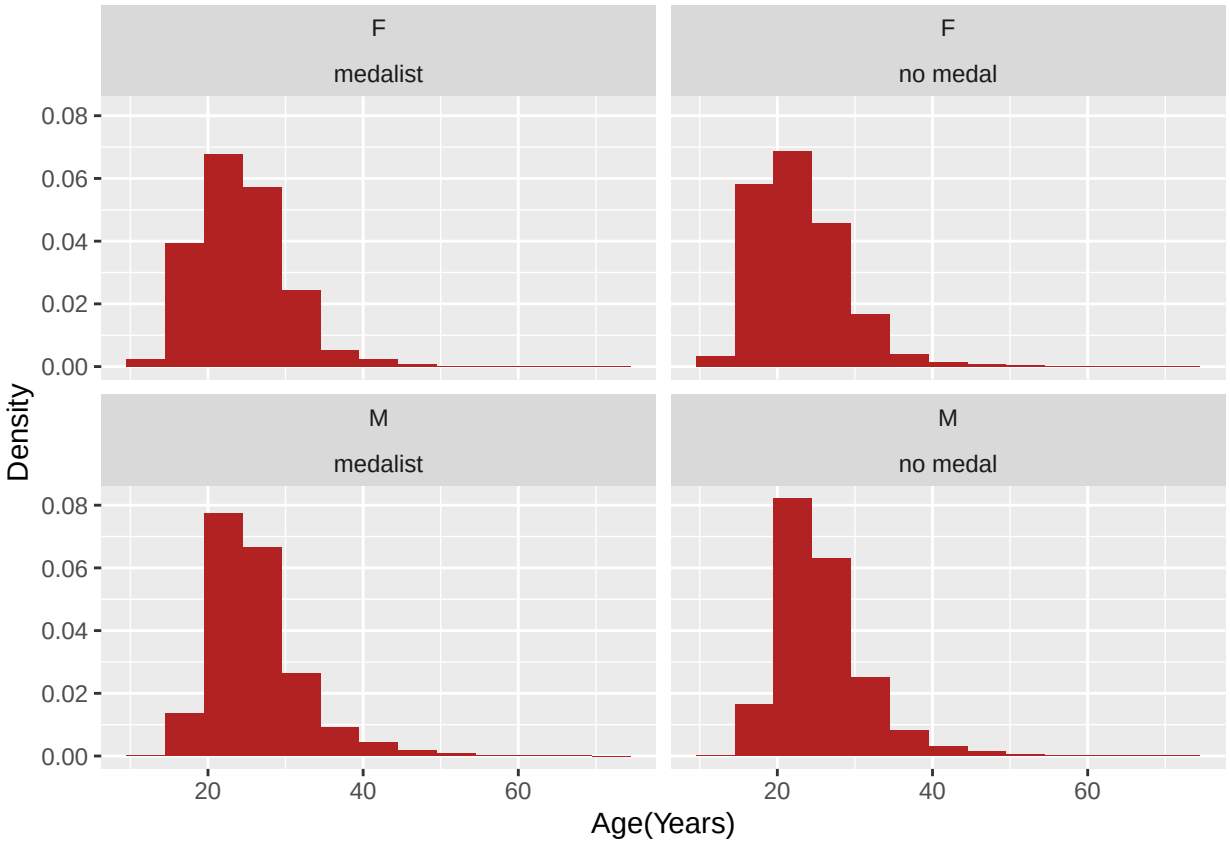
```
## Warning: Removed 2421 rows containing non-finite values (stat_ydensity).
```




```
# Generation of histogram subplots to depict athlete age densitites across each category: 1) Female med

ggplot(olympics_top, aes(age, y = after_stat(density))) +
  geom_histogram(binwidth = 5.0, center = 2,fill = "firebrick") +
  facet_wrap(vars(sex, medalist)) + xlab("Age(Years)") + ylab("Density")
```

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




**Discussion:** *In referencing the Violin plots in figure 1, it is shown that the vast majority of sports in the olympics feature athletes whose average age lies in the range of low to mid twenties. This is reasonable given that humans tend to reach physical and mental maturity (and thus optimal performing ability) around this time. Therefore, although there are some groups whose average age are higher or lower than others, the differences are quite marginal.Looking more closely at the plot, we see that the age distributions of the sports of basketball, boxing, and biathlon have the smallest spreads. For basketball and boxing, this is because the two sports are physical contact/impact sports where serious injuries are relatively common. These types of injuries (like ACL/MCL tears, concussions, etc.) are often difficult to rehabilitate and can be career ending for many athletes, hence, reducing their competition age. The biathlon sport is arguably the most all-around difficult sport to train and compete in because it requires athletes to be skilled in many areas including speed, strength, endurance, and precision. Additionally, the sport is very time-intensive as it encompasses many sub-events in a single race. It is then reasonable to conclude that the spread of the age distribution of biathlon athletes is low relative to other sports because it is difficult for an athlete to compete at a high-level in a sport that has a large number of components for a long period of time.*

*Keeping our discussion on figure 1, it is shown that equestrianism and archery have the two largest spreads. A possible explanation for why there are many athletes in these sports that participate over a wide age range is that these sports are predominantly measures of an athletes precision and accuracy (stemming from mental focus and strength) in a competition as opposed to his/her physical ability. Whereas an athlete's physical abilities tend to drop off rather sharply with time, an athlete often retains his/her mental abilities for decades after experiencing a drop-off in physical abilities. Therefore, it is sensible that the spread of equestrian and archery athlete ages is very large.*

*Shifting gears to the second figure, we have four histogram density subplots that visualize the distribution of athlete age according to sex, and whether or not they received a medal. In comparing the distributions, it is determined that the trends are very similar. The shape of the distributions are all unimodal and skewed at their upper-end. For each of the four categories, the highest age frequencies occur in the range of the low-mid*

*twenties as was the case for the trends depicted in figure 1 of the Part 2 Discussion. Now, although the trends are very similar, there are some slight differences in the distributions. In particular, there appears to be a noticeably steeper decline in the age frequencies from the mode for male medalists and male non-medalists than for female medalists, and female non-medalists.*

*A possible explanation for this trend is that the male sporting competitions tend to be more physically-demanding and contact-oriented than female sporting competitions. This is largely due to men, on average, having elevated physical abilities relative to their female counterparts (measured by average muscle tone and quality). With increased physical requirements in a sport comes increased potential for serious (potentially career-ending) injuries. Hence, it is reasonable to conclude that the proportion of female athletes that compete across each age group is more well-balanced than that of male athletes. In conclusion, although there are minor features that distinguish the histogram age distributions in figure 2, at a high-level, there are no major changes in the trends across the categories.*