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:

The dataset `olympics_top` *is a historical dataset containing information on Olympic games competitors from 1896 through Summer 2016. Variables include information on age, weight, height, sex, where the games were held, what event the athlete competed in, if they received a medal, and so on. This information is encoded long, with each observation encoding information about a particular athlete's performance in a particular Olympic games. Data was scraped from www.sports-references.com by a Kaggle user, after original compilation by sports hobbyists.*

To answer the question above, I'll need the information on sport, sex, height, and medalist status.

```
head(olympics_top)
```

```
## # A tibble: 6 x 18
##      id name     sex     age height weight team   noc   games  year season city
##   <dbl> <chr>    <chr> <dbl>  <dbl>  <dbl> <chr>  <chr> <chr> <dbl> <chr>  <chr>
## 1     5 Christi~ F        21    185     82 Nethe~ NED   1988~  1988 Winter Calg~
## 2     5 Christi~ F        21    185     82 Nethe~ NED   1988~  1988 Winter Calg~
## 3     5 Christi~ F        25    185     82 Nethe~ NED   1992~  1992 Winter Albe~
## 4     5 Christi~ F        25    185     82 Nethe~ NED   1992~  1992 Winter Albe~
## 5     5 Christi~ F        27    185     82 Nethe~ NED   1994~  1994 Winter Lill~
## 6     5 Christi~ F        27    185     82 Nethe~ NED   1994~  1994 Winter Lill~
## # ... with 6 more variables: sport <chr>, event <chr>, medal <chr>, gold <chr>,
## #   medalist <chr>, decade <dbl>
```

```
# Lets see what's in the data...
table(olympics_top$games)
```

```
##
## 1896 Summer 1900 Summer 1904 Summer 1906 Summer 1908 Summer 1912 Summer
##         127         708         665         388        1128         978
## 1920 Summer 1924 Summer 1924 Winter 1928 Summer 1928 Winter 1932 Summer
##        1225        1745         195        1709         262         846
## 1932 Winter 1936 Summer 1936 Winter 1948 Summer 1948 Winter 1952 Summer
##         193        2063         395        1729         527        2943
## 1952 Winter 1956 Summer 1956 Winter 1960 Summer 1960 Winter 1964 Summer
##         574        2035         709        2941         758        2926
## 1964 Winter 1968 Summer 1968 Winter 1972 Summer 1972 Winter 1976 Summer
##         948        2753         835        3283         793        3042
## 1976 Winter 1980 Summer 1980 Winter 1984 Summer 1984 Winter 1988 Summer
##         788        1879         782        2612         819        3300
## 1988 Winter 1992 Summer 1992 Winter 1994 Winter 1996 Summer 1998 Winter
##        1110        3465        1363        1143        3358        1219
## 2000 Summer 2002 Winter 2004 Summer 2006 Winter 2008 Summer 2010 Winter
##        3179        1291        3106        1428        3213        1383
## 2012 Summer 2014 Winter 2016 Summer
##        3098        1512        3055
```

```
# games is encoded as a character string, so we're going to need to generate a table to get an idea abo
```

Approach:

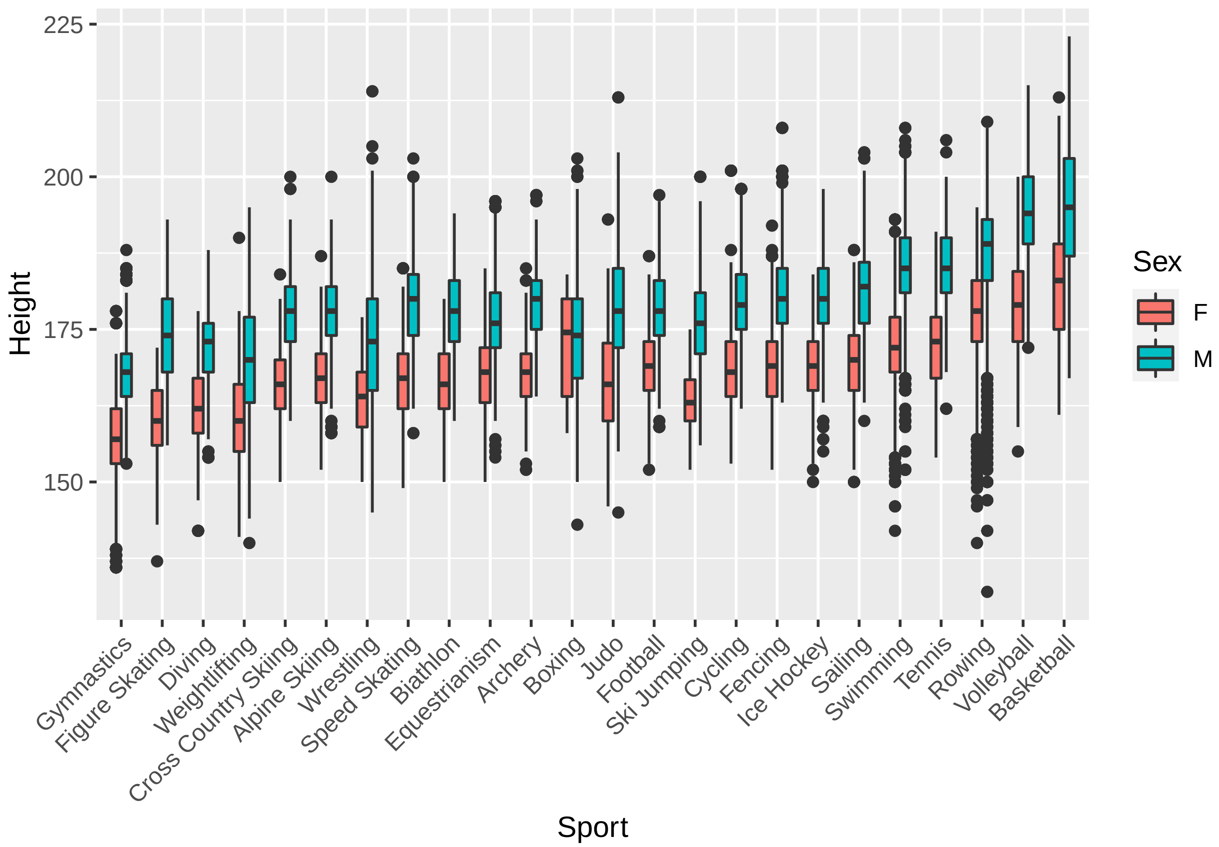
I'll answer the first part of the question by using the `geom_boxplot()` *feature of* `ggplot()` *to visualize height grouped by sport and sex. For the second part of the question, I'll use a* `geom_density_ridges()` *plot to visualize the distributions of height among medalists and non-medalists.*

Analysis:

```
#part one, boxplot()

ggplot(olympics_top, aes(reorder(sport, height, na.rm = TRUE), height, fill = sex)) +
  geom_boxplot(position = position_dodge(width = .5), width = .5) +
  scale_x_discrete(name = "Sport") +
  scale_y_continuous(name = "Height") +
  labs(fill = "Sex") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1, vjust = 1))
```

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

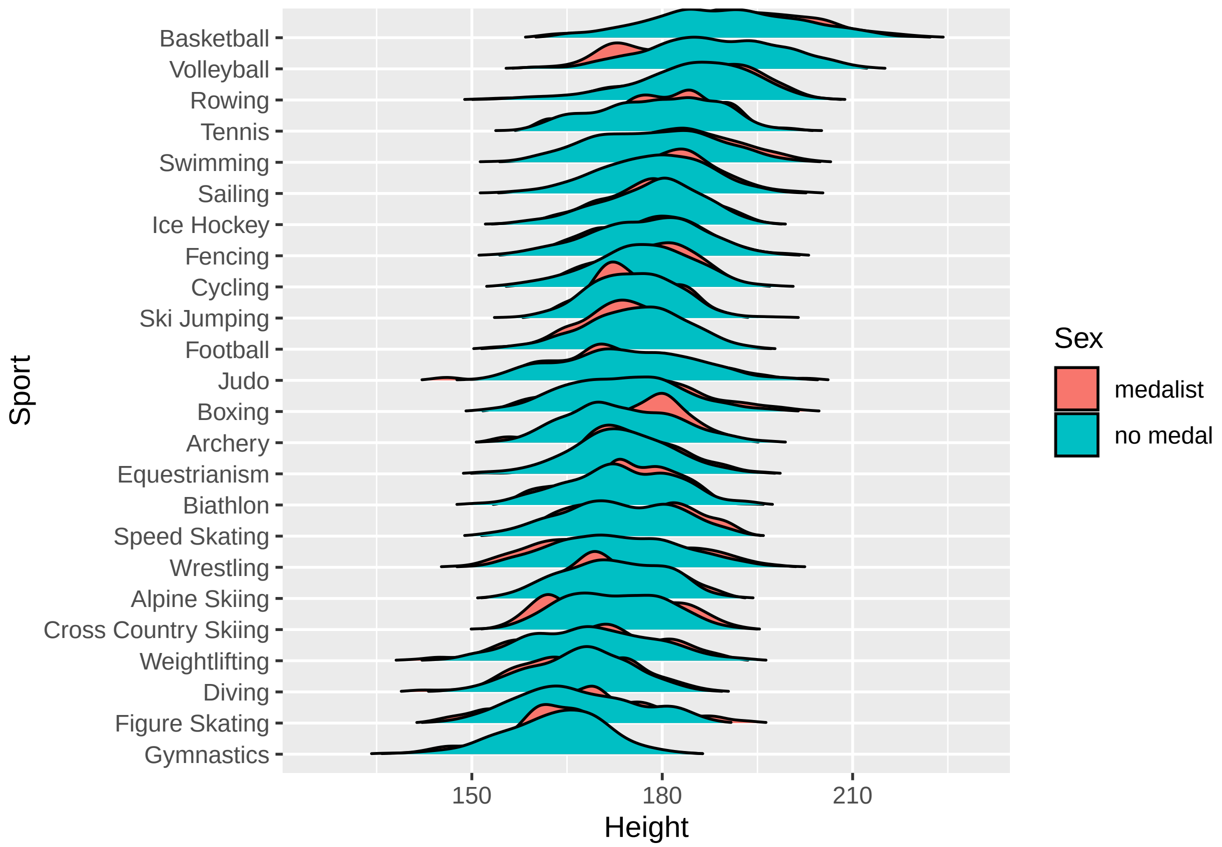


```
# part two, ridgeline()

ggplot(olympics_top, aes(height, reorder(sport, height, na.rm = TRUE), fill = medalist)) +
  geom_density_ridges(rel_min_height = 0.01) +
  scale_x_continuous(name = "Height") +
  scale_y_discrete(name = "Sport") +
  labs(fill = "Sex")
```

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

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



Discussion:

Gymnastics, figure skating, diving, and weightlifting all feature the shortest athletes. This is likely because the shorter athletes experience less joint strain, displace a smaller volume, and have better leverages respectively. Basketball, volleyball, rowing, and tennis have the tallest athletes. This is likely because athletes competing in these sports may benefit from longer limbs to exert greater leverage and extend their reach.

Medalists in gymnastics, all skiing types, wrestling, sailing, and fencing were shorter than the average competitor. Medalists in boxing, archery, equestrianism,swimming, tennis, rowing, basketball, and biathlon appear to be taller on average.

Part 2

Question:

Did communist/former communist regimes regularly produce better weightlifters than others? Do weightlifting athletes compete at the upper limits of their weight classes?

Introduction:

Soviet bloc and communist regimes were once famous for producing star strength athletes. I will use information on country, date, and medalist status to determine if Soviet/communist training styles were responsible for more medals/athlete than Western training regimes during the cold war. I will then use the information on weight and medalist status among weightlifters to test if weightlifters are more likely to compete at the upper limit for their weight class.

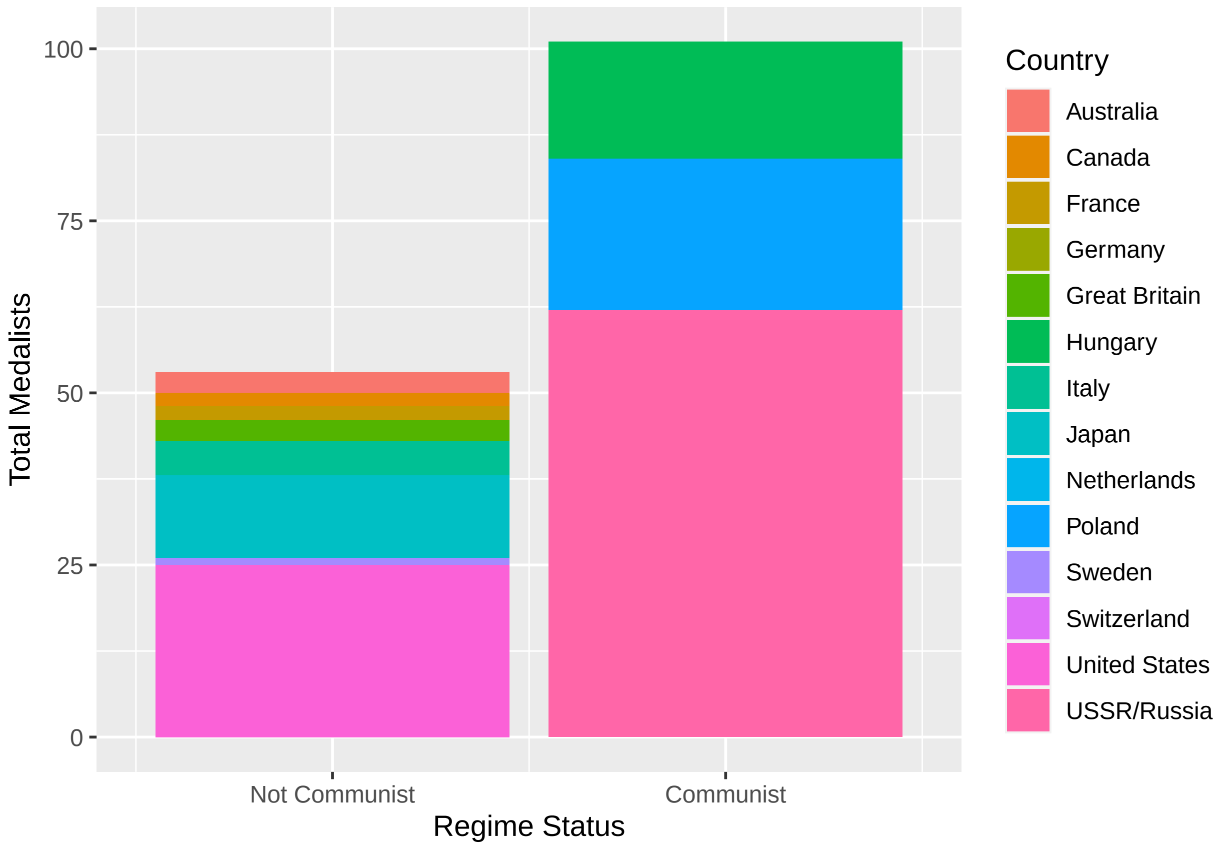
```
#dummy variable for communist regimes
olympics_top$commies <- ifelse(olympics_top$team == "USSR/Russia", 1, ifelse(olympics_top$team == "Pola
#dummy variable for medalist status
olympics_top$medal_count <- ifelse(olympics_top$medalist == "medalist", 1, ifelse(olympics_top$medalist
#this country data seemed really limited...
#filtering to cold war
olympics_top <- filter(olympics_top, year > 1948)
olympics_top <- filter(olympics_top, year <= 1991)
#filtering the data to only weightlifters
olympics_liftr <- filter(olympics_top, sport == "Weightlifting")
```

Approach:

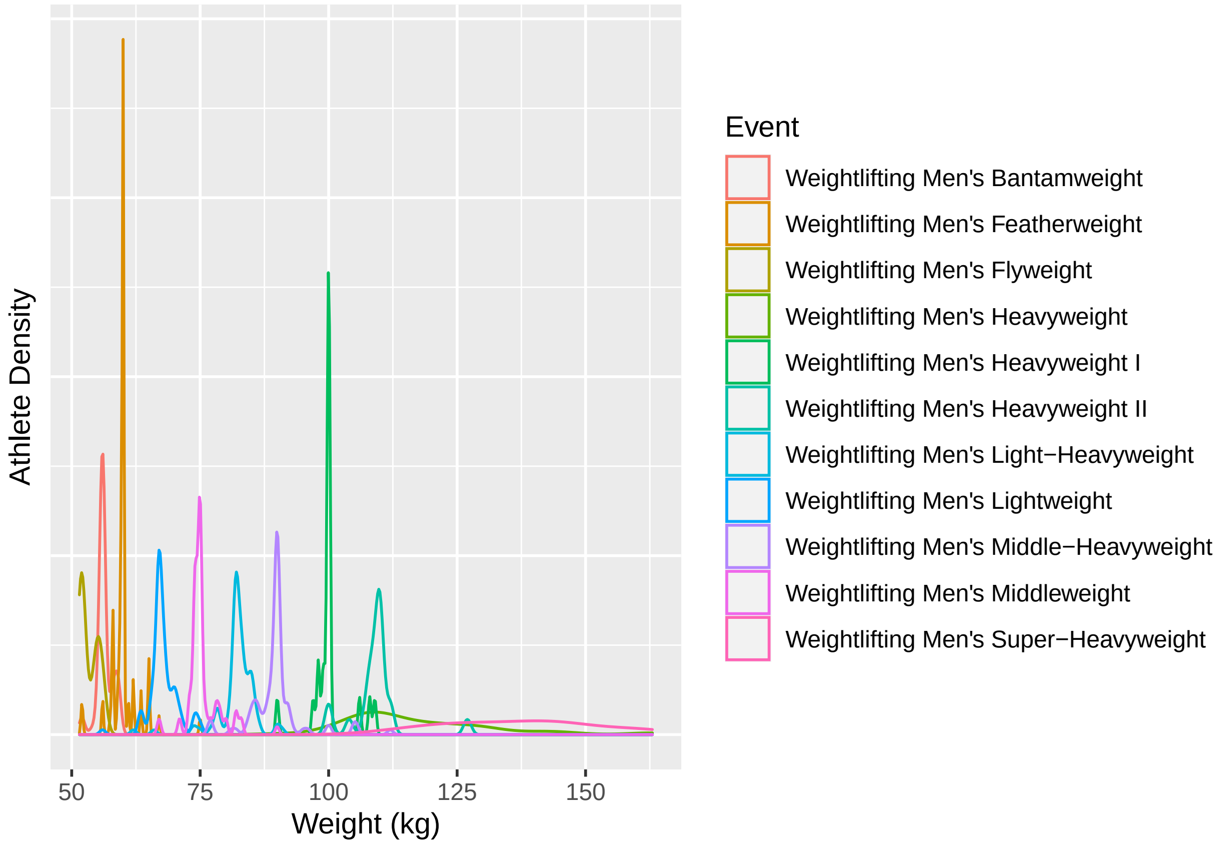
For the relationship between communist regime status and weightlifting performance, I will use a `geom_col()` *chart of total medalists sorted by communism status. For the weight distributions of weightlifters, I will use a* `geom_density()` *graph to see if spikes are obervable at the upper limit of weight classes.*

Analysis:

```
#geom_col for part 1
ggplot(olympics_liftr, aes(commies, medal_count, fill = team)) +
  geom_col() +
  labs(fill = "Country") +
  scale_x_continuous(name = "Regime Status",
                     breaks = c(0, 1),
                     limits = c(-0.5, 1.5),
                     labels = c("Not Communist", "Communist")) +
  scale_y_continuous(name = "Total Medalists")
```



```r
# geom_line and geom_jitter for part 2
ggplot(olympics_liftr, aes(weight, color = event)) +
  geom_density() +
  scale_x_continuous(name = "Weight (kg)") +
  labs(color = "Event", y = "Athlete Density") +
  theme(axis.text.y = element_blank(),
        axis.ticks.y = element_blank())
```



Discussion:

For the bar graph of communist/non-communist medalists in weightlifting, it is abundantly clear that communist countries produced more medalists. Poland alone produced almost as many medalist weightlifters as the United States, with the Soviet Union taking home more than twice the number of medals as the Americans. This was likely due to the abundant supply of trainees, widespread doping, and rigorous training regimens employed by Communist regimes during this period.

The density plot shows some pretty strong clustering at certain peaks, usually towards the upper limit beyond which the athletes will be unable to compete in the given weight class. This demonstrates the importance of power to weight ratio in olympic weightlifting, similar to the importance of weigh-ins to fighting sports.