Data Visualization: Winter Olympic Games Analysis Project

Ivan Meng

Introduction

In this project, we choose to discuss the data from the recent Winter Olympic Games which happened in China. For this project, we selected 3 specific datasets, where one contains information about athletes, one contains information about coaches, and one contains information about medals. In the athletes dataset, it contains athletes' personal information such as name, birthplace and country they are playing for. The coached dataset also present the same format of data for the coaches. In the medal dataset, the type of medals are specified and it also contains the athlete's name and country for each metal.

These datasets are acquire from the website 'kaggle' through the link https://www.kaggle.com/datasets/piterfm/beijing-2022-olympics?select=medals.csv (https://www.kaggle.com/datasets/piterfm/beijing-2022-olympics?select=medals.csv). They are interesting to our group because this event happened in our home country China and Zixiang also has similar personal experience playing sports within competition.

We expect to find a trend between the number of athletes from different countries and the total medals they acquire during the events. We also expect a relationship between the number of coaches and the total number of medals for that specific country.

Import the raw data and load packages

```
# Import the datasets from the computer
library(tidyverse)
athletes <-read.csv("athletes.csv")
coaches <-read.csv("coaches.csv")
medals <-read.csv("medals.csv")
```

DataFrame Manipulation

Use different functions such as unite() and seperate() to adjust the datasets into optimal format.

```
# Use separate() to pull apart height_m.ft into two columns
# Use unite() to rejoin the residence_place and residence_country
# Use rename() to set new names as labels
athletes <- athletes %>%
  separate(height_m.ft, into = c("height_meter", "height_feet"), sep = "/", convert = TR
UE, remove = FALSE, fill = "right") %>%
  unite(residence, residence_place, residence_country, sep = ", ") %>%
  rename(athlete_name = name) %>%
  rename(athlete_short_name = short_name) %>%
  rename(athlete_gender = gender) %>%
  rename(athlete_residence = residence) %>%
  rename(athlete_height_meter = height_meter) %>%
  rename(athlete_height_feet = height_feet) %>%
  rename(athlete_birth_date = birth_date) %>%
  rename(athlete_birth_place = birth_place) %>%
  rename(athlete_birth_country = birth_country) %>%
  rename(athlete_link = url) %>%
  select(-height_m.ft)
```

Reformat the medal_date

```
# change the date format to get rid off the extra data on the date column
medals$medal_date <- as.Date(medals$medal_date, format = "%Y-%m-%d")
medals <- medals %>% arrange(athlete_name)
```

Rename the datasets “coaches” to simplfy the data

```
# reorder the column "name" by alphabetical order
coaches <- coaches %>%
  rename(coach_name = name) %>%
  rename(coach_gender = gender) %>%
  rename(coach_birth_date = birth_date) %>%
  rename(athlete_link = url) %>%
  select (-event)
```

Joining/Merging

Take a first look at the information of each dataset

```
# take a look at the dataset "athlete" with glimpse
glimpse(athletes)
```

```
## Rows: 2,897
## Columns: 14
## $ athlete_name          <chr> "AAGAARD Mikkel", "AALTO Antti", "AALTONEN Miro"…
## $ athlete_short_name    <chr> "AAGAARD M", "AALTO A", "AALTONEN M", "ABDELKADE…
## $ athlete_gender        <chr> "Male", "Male", "Male", "Male", "Male", "Male", …
## $ athlete_birth_date    <chr> "1995-10-18", "1995-04-02", "1993-06-07", "1987-…
## $ athlete_birth_place   <chr> "FREDERIKSHAVN", "KITEE", "JOENSUU", "MUSKEGON, …
## $ athlete_birth_country <chr> "Denmark", "Finland", "Finland", "United States …
## $ country               <chr> "Denmark", "Finland", "Finland", "United States …
## $ country_code          <chr> "DEN", "FIN", "FIN", "USA", "KSA", "USA", "ERI",…
## $ discipline            <chr> "Ice Hockey", "Ski Jumping", "Ice Hockey", "Ice …
## $ discipline_code       <chr> "IHO", "SJP", "IHO", "IHO", "ALP", "BOB", "ALP",…
## $ athlete_residence     <chr> "ORNSKOLDSVIK, Sweden", "KUOPIO, Finland", "PODO…
## $ athlete_height_meter  <dbl> 1.84, NA, 1.80, 1.87, NA, NA, NA, 1.95, 1.93, NA…
## $ athlete_height_feet   <chr> "6'0''", NA, "5'10''", "6'1''", NA, NA, NA, "6'4…
## $ athlete_link          <chr> "../../../en/results/ice-hockey/athlete-profile-…
```

```
# take a look at the dataset "coaches" with glimpse
glimpse(coaches)
```

```
## Rows: 77
## Columns: 7
## $ coach_name       <chr> "BARES Jakub", "BEIGHTON Sean", "BEIGHTON Sean", "BEL…
## $ coach_gender     <chr> "Male", "Male", "Male", "Male", "Male", "Male", "Male…
## $ coach_birth_date <chr> "1988-03-20", "1988-11-22", "1988-11-22", "1961-08-04…
## $ country          <chr> "Czech Republic", "United States of America", "United…
## $ discipline       <chr> "Curling", "Curling", "Curling", "Curling", "Ice Hock…
## $ function.        <chr> "Coach", "Coach", "Coach", "Coach", "Head Coach", "Co…
## $ athlete_link     <chr> "../../../en/results/curling/athlete-profile-n1034345…
```

```
# dataset "medals"
glimpse(medals)
```

```
## Rows: 694
## Columns: 12
## $ medal_type         <chr> "Gold", "Silver", "Silver", "Silver", "Silver", "Go…
## $ medal_code         <int> 1, 2, 2, 2, 2, 1, 1, 3, 2, 1, 1, 3, 1, 1, 2, 2, 2, …
## $ medal_date         <date> 2022-02-20, 2022-02-16, 2022-02-20, 2022-02-15, 20…
## $ athlete_short_name <chr> "AALTONEN M", "ABRAMENKO O", "AICHER E", "ALDOSHKIN…
## $ athlete_name       <chr> "AALTONEN Miro", "ABRAMENKO Oleksandr", "AICHER Emm…
## $ athlete_sex        <chr> "M", "M", "X", "M", "W", "W", "M", "W", "M", "W", "…
## $ athlete_link       <chr> "../../../en/results/ice-hockey/athlete-profile-n10…
## $ event              <chr> "Men", "Men's Aerials", "Mixed Team Parallel", "Men…
## $ country            <chr> "Finland", "Ukraine", "Germany", "ROC", "Germany", …
## $ country_code       <chr> "FIN", "UKR", "GER", "ROC", "GER", "CAN", "NOR", "S…
## $ discipline         <chr> "Ice Hockey", "Freestyle Skiing", "Alpine Skiing", …
## $ discipline_code    <chr> "IHO", "FRS", "ALP", "SSK", "SJP", "IHO", "NCB", "C…
```

The dataset “athletes” contains 2897 rows/observations and 14 columns/variables. The dataset “athletes” has 7 unique IDs: “athlete_birth_date”, “athlete_birth_place”, “athlete_birth_country”, “athlete_residence”, “residence_country”, athlete_height_meter”, and “athlete_height_feet”. The dataset “athletes” with dataset “medals” have 7 common IDs: “athlete_name”, “athlete_short_name”, “country”, “country_code”, “discipline”, “discipline_code”, and “athlete_link”.

The dataset “medals” contains 694 rows/observations and 12 columns/variables. It has 5 unique IDs are “medal_type”, “medal_code”, “medal_date”, “athlete_sex”, and “event”. The dataset “medals” with atheltes “medals” have 7 common IDs: “athlete_short_name”, “athlete_name”, “athlete_link”, “country”, “country_code”, “discipline”, and “discipline_code”.

The dataset “coaches” contains 77 rows/observations and 7 columns/variables. Its unique IDs are “coach_name”, “coach_gender”, “coach_birth_date”, “event”, and “function”. Its common IDs are “country”, “discipline”, and “athlete_link”.

Now we want to join the data of athletes with medals to have the athletes matched with their medals, then we want to join that dataset with coaches to have each coache matched with their corresponding dicipline and country.

```
# Join the separated data sources into combined dataset
# Join "medals" into "athletes" by using right_join() and common variables: "athlete_name", "athlete_short_name", "country", "country_code", "discipline", "discipline_code", "athlete_link" and delete the column "athlete_sex".
athletes_with_medals <- athletes %>%
  right_join(medals, by = c("athlete_name", "athlete_short_name", "country", "country_code", "discipline", "discipline_code", "athlete_link")) %>%
  select(- athlete_sex)

#check if there are any duplicates
sum(duplicated(athletes_with_medals))
```

```
## [1] 0
```

```
# Join dataset "coaches" into the new dataset "athletes_with_medals" by using left_join and common variables" "country" and "discipline".
athletes_with_medals_coaches <- athletes_with_medals %>%
  left_join(coaches, by = c("country", "discipline"))
```

I combine the datasets “athletes” and “medals” as a new dataset “athletes_with_medals”, which contians 694 rows/observations. There are 2897 athletes participate the tournament and 694 athletes win medals. Therefore, there are 2203 rows/observations dropped when joinning the two datasets.

I combine the dataset “coaches” and the “athletes_with_medals” as final dataset “athletes_with_medals_coaches”, which contians 986 rows/observations. Some athletes have more than one coaches. Therefore, there are 292 rows/observations added when joining the two datasets.

The related potential issues is that the athletes may have duplicated coaches. However, the relationship between coaches and medals are weak. We can not find the useful insight from the data.

Wrangling

```
# first we can find out the proportion of male athletes who won a medal
athletes_with_medals %>%
  select(athlete_gender) %>%
  summarize(proportions_male = mean(athlete_gender == "Male", na.rm = T))
```

```
##   proportions_male
## 1        0.5508721
```

```
# then we can find out the proportion of female athletes who won a medal the same way
athletes_with_medals %>%
  select(athlete_gender) %>%
  summarize(proportions_female = mean(athlete_gender == "Female", na.rm = T))
```

```
##   proportions_female
## 1          0.4491279
```

As we see above, in this winter Olympics, the proportion of males who won a medal is about 55%, where the female proportion is about 45%. The proportion of male is about 10 percent higher than that of females.

Even though heights may have different effects in different games, we still want to discuss if this higher proportion of medal winning rate related to gender has anything to do with the average height of this gender

```
# calculate the mean height of the male athletes
athletes_with_medals %>%
  filter(athlete_gender == "Male") %>%
  summarise(mean_height = mean(athlete_height_meter,na.rm = TRUE))
```

```
##   mean_height
## 1       1.842
```

```
# # calculate the mean height of the female athletes
athletes_with_medals %>%
  filter(athlete_gender == "Female") %>%
  summarise(mean_height = mean(athlete_height_meter,na.rm = TRUE))
```

```
##   mean_height
## 1    1.696667
```

The mean height for male is 1.842 meters, where the mean height of female is 1.697 meters. We can see that in general, the male athletes are higher compare to the female athletes.

But is height really a factor for winning a medal? We decide to calculate the average height for each type of medals (for example, gold, silver and etc). (This is not quite precise because for each type of game, different height range is optimal, but it is still interesting to see if it has a effect)

```
# calculate the mean height for athletes who on a gold medal
athletes_with_medals %>%
  filter(medal_type == "Gold") %>%
  summarise(mean_height = mean(athlete_height_meter,na.rm = TRUE))
```

```
##   mean_height
## 1    1.782206
```

```
# calculate the mean height for athletes who on a silver medal
athletes_with_medals %>%
  filter(medal_type == "Silver") %>%
  summarise(mean_height = mean(athlete_height_meter,na.rm = TRUE))
```

```
##   mean_height
## 1    1.775362
```

```
# calculate the mean height for athletes who on a bronze medal
athletes_with_medals %>%
  filter(medal_type == "Bronze") %>%
  summarise(mean_height = mean(athlete_height_meter,na.rm = TRUE))
```

```
##   mean_height
## 1    1.778158
```

As shown above, even though the mean height for the gold medal winners are slightly higher than the other two types, they are not drastically different from each others. Thus, we concluded that generally, height is not a critical factor for winning a medal in the winter Olympic games.

Since both of us come from China, we are curious about what proportion of athletes who won medals comes from China

```
# calculate the proportion of athletes who come from China
athletes_with_medals %>%
  select(country) %>%
  summarize(proportions = mean(country == "People's Republic of China"))
```

```
##   proportions
## 1  0.03746398
```

As calculated, about 3.7 percent of athletes who on a medal in the winter Olympic game did come from China.

One of our group’s members hometown is Harbin, China. We want to further investigate if there are any athletes who come from Harbin who won a medal in the winter Olympics. And if so, how many?

```
# calculate the total number of athletes who won a medal in the game and comes from Harb
in
athletes_with_medals %>%
  filter(athlete_birth_place == "HARBIN") %>%
  summarize(athletes_from_Harbin=n())
```

```
##   athletes_from_Harbin
## 1                    5
```

```
# calculate the percentage of athletes from Harbin
athletes_with_medals %>%
  mutate(if_harbin = ifelse(athlete_birth_place=='HARBIN',T,F)) %>%
  select(if_harbin) %>%
  summarise_all(mean)
```

```
##   if_harbin
## 1        NA
```

Surprisingly, there are 5 athletes who came from Harbin and won a medal in the winter Olympics. This is about 0.7 percent of the entire group of the athletes.

Now we want to discuss whether the number of athletes has a effect on the medals counts for a specific country. Thus we joined the data accordingly.

```
# summaries the total medal counts for each country
medalcount <- athletes_with_medals %>%
  group_by(country) %>%
  summarise(total_medals=n())

# summaries the total athletes counts for each country
athletescount <- athletes %>%
  group_by(country) %>%
  summarise(total_athletes = n())

# joined the two datasets to have a complimentry view of how the number of athletes affe
cts the medal counts
athletes_medals_percountry <- athletescount %>%
  left_join(medalcount, by = c("country"))%>%
  mutate_all(~replace(., is.na(.), 0)) %>%
  arrange(desc(total_medals))

# check whether the countries is distinct
athletes %>% summarize(n_distinct(country))
```

```
##   n_distinct(country)
## 1                  91
```

We are also interested in whether the number of coaches has a effect on the medal counts for a country, so we do the same for the coaches dataset.

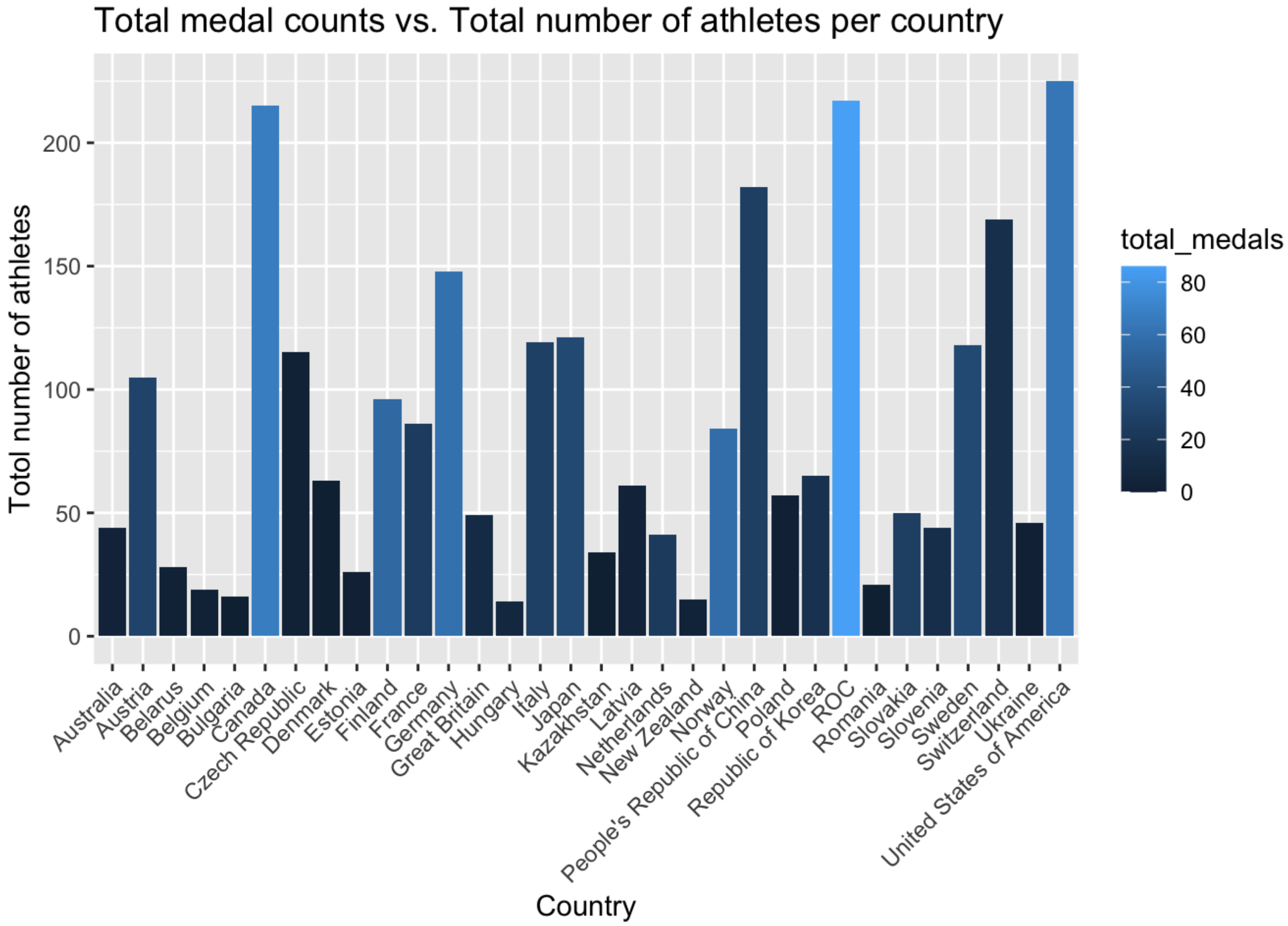
```
# find out the to total number of coaches and total number of medals
coach_medal <- athletes_with_medals_coaches %>%
  group_by(country) %>%
  summarise(total_coaches = n_distinct(coach_name), total_medals = n_distinct(athlete_na
me)) %>%
  arrange(desc(total_medals))

head(coach_medal)
```

```
## # A tibble: 6 × 3
##   country                  total_coaches total_medals
##   <chr>                            <int>        <int>
## 1 ROC                                  3           68
## 2 Canada                               8           59
## 3 United States of America             3           57
## 4 Finland                              4           52
## 5 Germany                              1           47
## 6 Norway                               4           34
```

Creating Visualization

To get a general visualization of the relationship, we used a bar plots to see whether the number of athletes has a effect on the medal counts for the top 32 countries on the medal counts.

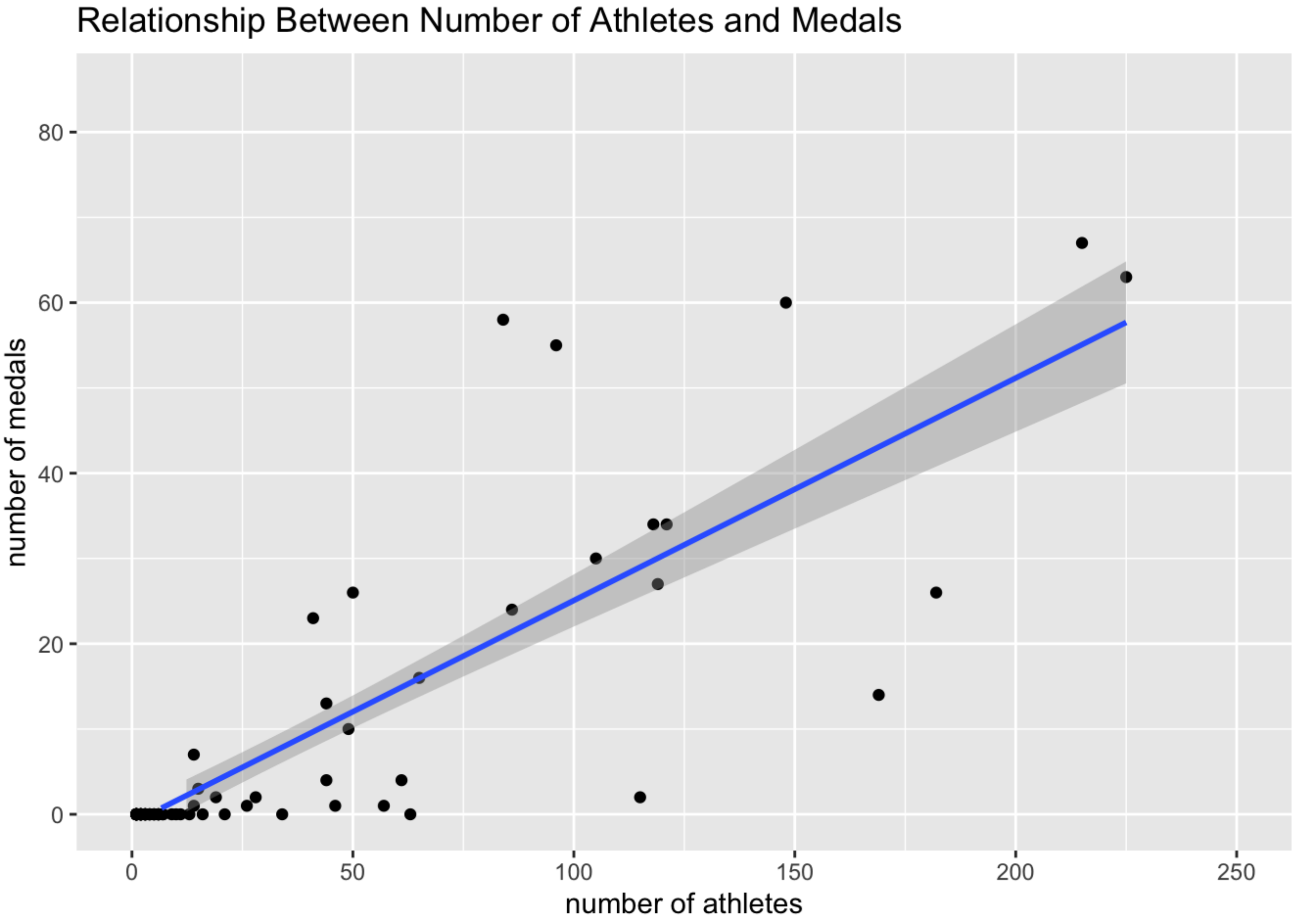
```
# create a plot showing the relationship between total medal counts and number of athlet
es
athletes_medals_percountry %>%
   arrange(desc(total_athletes))%>%
   slice(1:32)%>%
   ggplot(aes(x = country, y = total_athletes, fill = total_medals)) +
   geom_bar(stat="identity")+
   theme(axis.text.x = element_text(angle = 45, hjust = 1))+
   labs(title = "Total medal counts vs. Total number of athletes per country", x = "Coun
try", y = "Totol number of athletes")
```



In this graph, we can see that for countries with more athletes, they relatively have more medals earned in the Olympics. Countries with less than 50 athletes usually earned around 10 medals, where countries with more than 200 athletes earned more than 80 medals. Among all the countries, the US, Russia and Canada have the most athletes playing in the Olympic, which corresponded to a relatively high number of medals.

To further investigate the exact relationship, we used a point plot to calculate the effects of athletes' count on medals.

```
# Relationship between number of athletics and number of medals
ggplot(data = athletes_medals_percountry, aes(x=total_athletes, y=total_medals)) +
  geom_point() +
  geom_smooth(method='lm') +
  scale_x_continuous(n.breaks = 5) +
  labs(title = "Relationship Between Number of Athletes and Medals ") +
  scale_x_continuous("number of athletes", limits = c(0, 250)) +
  scale_y_continuous("number of medals", limits = c(0, 85))
```



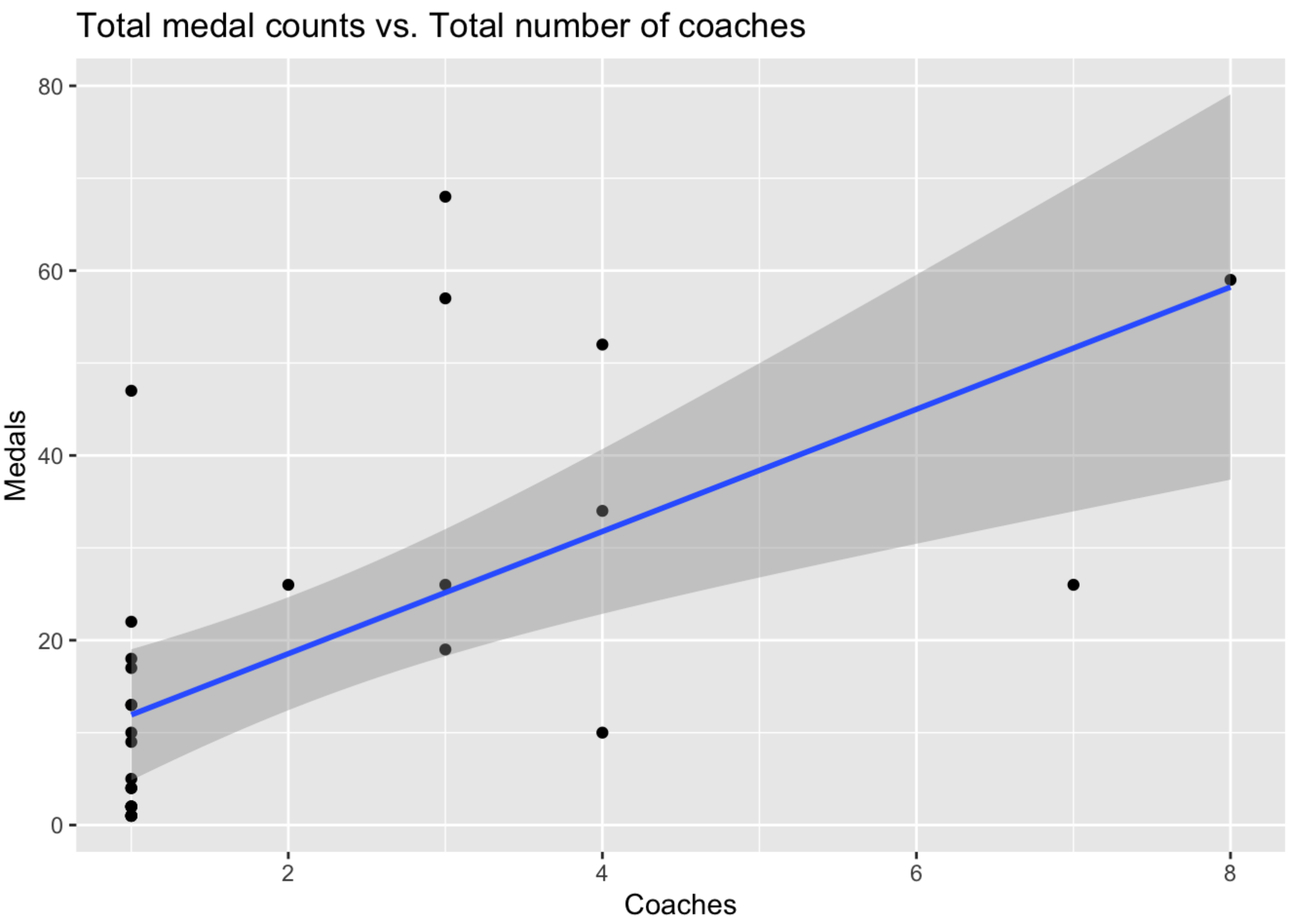
```
# investigate the numeric relatioship specifically
model <- lm(total_medals ~ total_athletes, data=athletes_medals_percountry)
summary(model)
```

```
##
## Call:
## lm(formula = total_medals ~ total_athletes, data = athletes_medals_percountry)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -32.393  -0.607   0.523   1.088  35.631
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -1.37106    1.08919  -1.259    0.211
## total_athletes  0.28263    0.01753  16.121   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.922 on 89 degrees of freedom
## Multiple R-squared:  0.7449, Adjusted R-squared:  0.742
## F-statistic: 259.9 on 1 and 89 DF,  p-value: < 2.2e-16
```

The graph shows the correlation between number of athletes and number of medals for countries with medals. When the country has one more athletes participate in the tournament, the number of medal of the country would increase 0.29 unit with standard error by 0.046. The predicted model would be Y = X(0.29) - 1.83.

Further on, we are curious if the number of coaches has a effect on the medal counts for each country. We decide to use a point plot and do a best fit to see if these two variables are related, just like what we did in the last part.

```
# make a plot that shows the relationship between number of coaches and number of medals
coach_medal %>%
  ggplot(aes(x = total_coaches, y = total_medals))+
  geom_point() +
  geom_smooth(method = "lm", formula= y~x) +
  labs(title = "Total medal counts vs. Total number of coaches", x = "Coaches", y = "Medals") +
  scale_x_continuous(n.breaks = 4 )
```



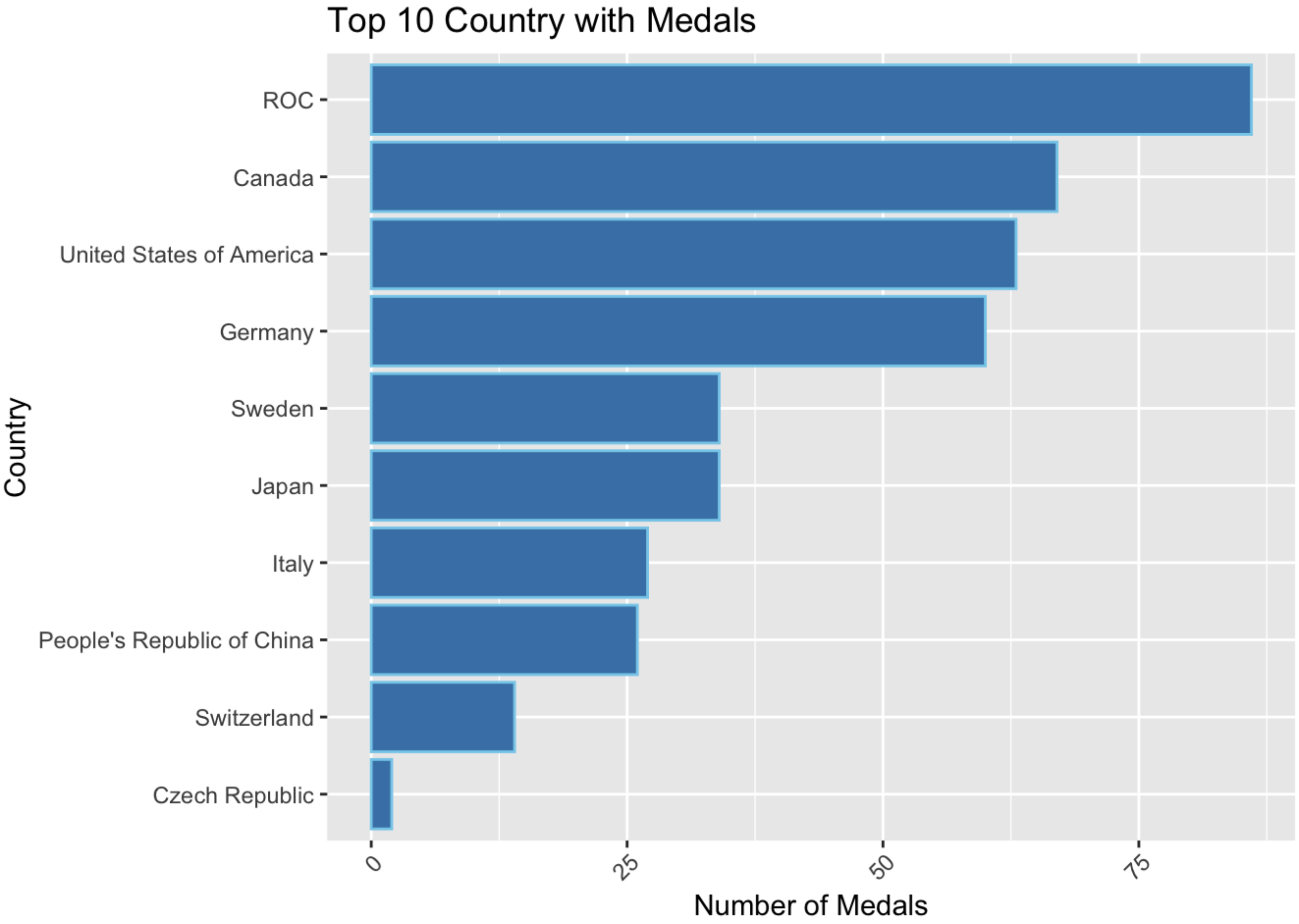
```
model2 <- lm(total_medals ~ total_coaches, data=coach_medal)
summary(model2)
```

```
## 
## Call:
## lm(formula = total_medals ~ total_coaches, data = coach_medal)
## 
## Residuals:
##    Min     1Q Median     3Q    Max 
## -25.61  -9.93  -2.93   5.07  42.84 
## 
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)    
## (Intercept)      5.317      4.511   1.179   0.2489    
## total_coaches    6.614      1.638   4.038   0.0004 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 16.04 on 27 degrees of freedom
## Multiple R-squared:  0.3765, Adjusted R-squared:  0.3534 
## F-statistic: 16.31 on 1 and 27 DF,  p-value: 0.0004001
```

In this graph, we can see that there is a weak relationship between the number of coaches and the number of total medals earn by a country. It does seems like more coaches would result in more medals for the country that the coaches came from, but there were still many countries who earned a lot of medals with only one coach. The relationship between these two variables are not very obvious.

The following graph show the Top 10 Country

```
# plot the graph
athletes_medals_percountry %>%
   arrange(desc(total_athletes))%>%
   slice(1:10)%>%
   ggplot(aes(x = reorder(country, total_medals), y = total_medals)) +
   geom_bar(stat="identity", color='skyblue', fill='steelblue')+
   coord_flip() +
   theme(axis.text.x=element_text(angle=45, hjust=0.9))+
   labs(title = "Top 10 Country with Medals", x = "Country", y =  "Number of Medals")
```


Top 10 Country with Medals
Country
ROC
Canada
United States of America
Germany
Sweden
Japan
Italy
People's Republic of China
Switzerland
Czech Republic
0
25
50
75
Number of Medals

```
##
sysname
##
"Darwin"
##
release
##
"21.5.0"
##
version
## "Darwin Kernel Version 21.5.0: Tue Apr 26 21:08:22 PDT 2022; root:xnu-8020.121.3~4/RE
LEASE_X86_64"
##
nodename
##                                                                                "MacBo
ok-Pro.local"
##
machine
##
"x86_64"
##
login
##
"root"
##
user
##
"zixiangmeng"
##                                                                                     e
ffective_user
##
"zixiangmeng"
```