**R CODE FOR OLYMPICS.CSV**

Install.packages("readr")

library(readr)

data=read.csv(choose.files())

head(data)

str(data)

> 'data.frame':   15316 obs. of  11 variables:

$ City      : chr  "Montreal" "Montreal" "Montreal" "Montreal" ...

$ Year      : int  1976 1976 1976 1976 1976 1976 1976 1976 1976 1976 ...

$ Sport     : chr  "Aquatics" "Aquatics" "Aquatics" "Aquatics" ...

$ Discipline : chr  "Diving" "Diving" "Diving" "Diving" ...

$ Event     : chr  "3m springboard" "3m springboard" "3m springboard" "3m springboard" ...

$ Athlete    : chr  "K\xd6HLER, Christa" "KOSENKOV, Aleksandr" "BOGGS, Philip George" "CAGNOTTO, Giorgio Franco" ...

$ Gender    : chr  "Women" "Men" "Men" "Men" ...

$ Country_Code: chr  "GDR" "URS" "USA" "ITA" ...

$ Country    : chr  "East Germany" "Soviet Union" "United States" "Italy" ...

$ Event_gender: chr  "W" "M" "M" "M" ...

$ Medal     : chr  "Silver" "Bronze" "Gold" "Silver" ...

Summary(data)

City          Year      Sport       Discipline

Length:15316     Min.  :1976  Length:15316     Length:15316

Class :character   1st Qu.:1984  Class :character  Class :character

Mode  :character   Median :1996  Mode  :character  Mode  :character

            Mean   :1994

            3rd Qu.:2004

Max.   :2008

   Event          Athlete          Gender          Country_Code

Length:15316     Length:15316     Length:15316     Length:15316

Class :character  Class :character  Class :character  Class :character

Mode :character  Mode :character  Mode :character  Mode :character




   Country       Event_gender       Medal

Length:15316     Length:15316     Length:15316

Class :character  Class :character  Class :character

Mode :character  Mode :character  Mode :character


## convert chr into factor

Cols_to_factor=c("City", "Sport", "Discipline", "Event", "Gender", "Country_Code", "Country",  "Event_gender", "Medal")

data[cols_to_factor]=lapply(data[cols_to_factor},as.factor) (as.factor convert chr to factor)

str(data)

> 'data.frame':   15316 obs. of  11 variables:

$ City      : Factor w/ 9 levels "Athens","Atlanta",..: 6 6 6 6 6 6 6 6 6 6 ...

$ Year      : int  1976 1976 1976 1976 1976 1976 1976 1976 1976 1976 ...

$ Sport     : Factor w/ 28 levels "Aquatics","Archery",..: 1 1 1 1 1 1 1 1 1 1 ...

$ Discipline  : Factor w/ 41 levels "Archery","Artistic G.",..: 14 14 14 14 14 14 14 14 14 14 ...

$ Event     : Factor w/ 293 levels "\xe9p\xe9e individual",..: 62 62 62 62 44 44 44 62 44 44 ...

$ Athlete    : chr  "K\xd6HLER, Christa" "KOSENKOV, Aleksandr" "BOGGS, Philip George" "CAGNOTTO, Giorgio Franco" ...

$ Gender     : Factor w/ 2 levels "Men","Women": 2 1 1 1 2 1 2 2 1 1 ...

$ Country_Code: Factor w/ 128 levels "AFG","AHO","ALG",..: 41 120 122 57 122 122 120 122 57 120 ...

$ Country    : Factor w/ 127 levels "Afghanistan",..: 30 96 118 52 118 118 96 118 52 96 ...

$ Event_gender: Factor w/ 3 levels "M","W","X": 2 1 1 1 2 1 2 2 1 1 ...

$ Medal      : Factor w/ 3 levels "Bronze","Gold",..: 3 1 2 3 1 3 2 1 2 1 ...

Summary(data)

```
    City              Year              Sport             Discipline


Beijing :2042 Min. :1976 Aquatics :2210 Athletics :1523

 Sydney :2015 1st Qu.:1984 Athletics :1523 Swimming :1422

 Athens :1998 Median :1996 Rowing :1377 Rowing :1377

 Atlanta :1859 Mean :1994 Hockey : 817 Hockey : 817

 Barcelona:1705 3rd Qu.:2004 Gymnastics: 783 Handball : 780

 Seoul :1546 Max. :2008 Handball : 780 Artistic G.: 672

 (Other) :4151 (Other) :7826 (Other) :8725
```

Event Athlete Gender

hockey : 817 Length:15316 Men :9388

handball : 780 Class :character Women:5928

football : 669 Mode :character

volleyball : 647

basketball : 646

eight with coxswain (8+): 486

(Other) :11271

Country_Code Country Event_gender Medal

USA :1992 United States:1992 M:8817 Bronze:5258

URS :1021 Soviet Union :1021 W:5773 Gold :5042

AUS : 798 Australia : 798 X: 726 Silver:5016

GER : 691 Germany : 691

CHN : 679 China : 679

RUS : 638 Russia : 638

(Other):9497 (Other) :9497

#how many medals were won in each year

medals_per_year=table(data$Year)

medals_per_year

1976 1980 1984 1988 1992 1996 2000 2004 2008

1305 1387 1459 1546 1705 1859 2015 1998 2042

| 1976 | 1980 | 1984 | 1988 | 1992 | 1996 | 2000 | 2004 | 2008 |
|------|------|------|------|------|------|------|------|------|
| 1305 | 1387 | 1459 | 1546 | 1705 | 1859 | 2015 | 1998 | 2042 |

1976 1980 1984 1988 1992 1996 2000 2004 2008

1305 1387 1459 1546 1705 1859 2015 1998 2042

sorted_medals = sort(medals_per_year, descending =TRUE) sorted_medals

1976 1980 1984 1988 1992 1996 2004 2000 2008 1305 1387 1459 1546 1705 1859 1998 2015 2042

```
barplot(sorted_medals, main="Number of Medals per Olympic Year",xlab = "Year",ylab =
"Number of Medals",las = 2,col="pink",cex.names=0.8)
```

**Number of Medals per Olympic Year**
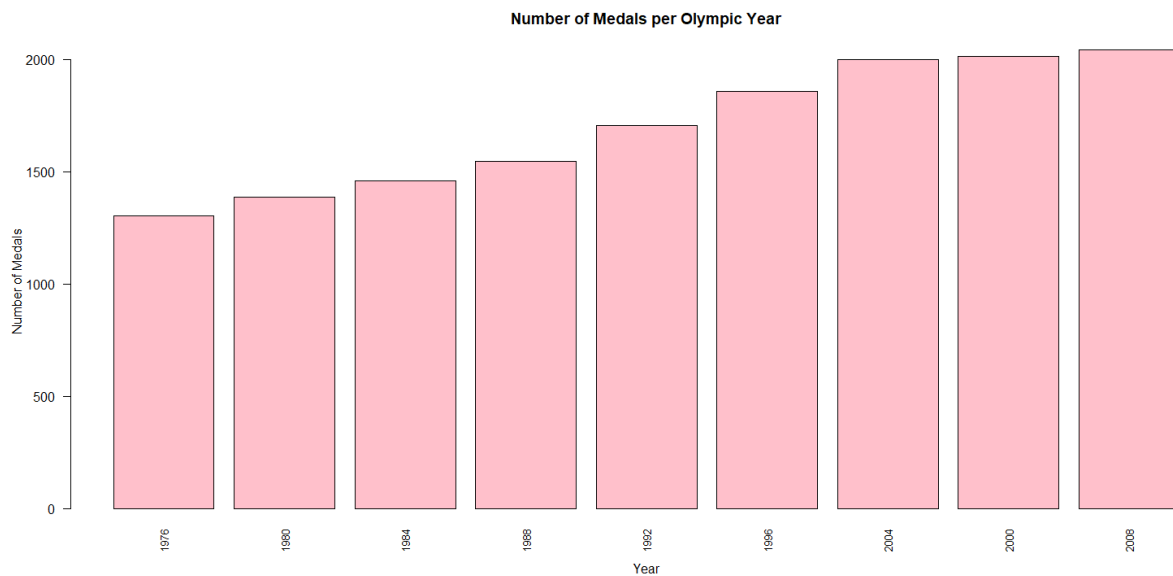


##are some sports becoming more popular over time

```
Library(dplyr)

df=data%>%select(Year,Sport)

sports_trend=df%>%group_by(Year,Sport)%>%

sports_trend=df%>%group_by(Year,Sport)%>%summarise(count=n(),.groups="drop")%>
%arrange(Sport,Year)

trend_by_sport=sports_trend%>%group_by(Sport)%>%summarise(growth=last(count)-
first(count))%>%arrange(desc(growth))

trend_by_sport

# A tibble: 28 × 2

Sport growth

 <fct> <int>

 1 Aquatics 188
```

2 Football 57

3 Hockey 50

4 Cycling 35

5 Judo 32

6 Athletics 30

7 Canoe / Kayak 27

8 Gymnastics 25

9 Shooting 24

10 Archery 18

11 Sailing 18

12 Weightlifting 18

13 Baseball 12

14 Volleyball 12

15 Wrestling 11

16 Taekwondo 8

17 Table Tennis 6

18 Equestrian 3

19 Handball 1

20 Badminton 0

21 Basketball 0

22 Boxing 0

23 Softball 0
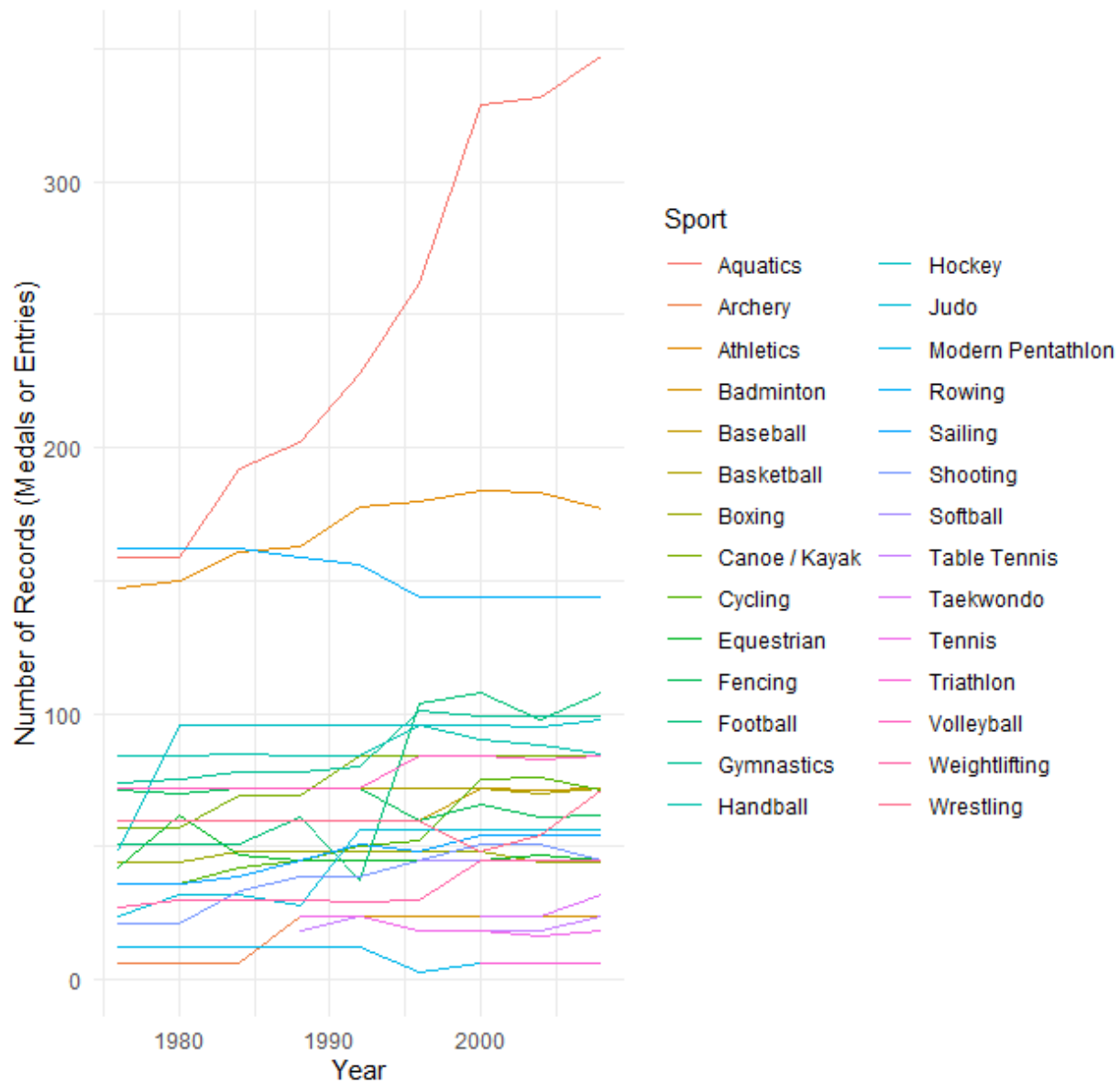
24 Triathlon 0

25 Modern Pentathlon -6

26 Tennis -6

27 Fencing –9

28 Rowing –18

library(ggplot2)

ggplot(sports_trend,aes(x=Year,y=count,color=Sport))+geom_line()+labs(title = "Are Some Sports Becoming More Popular Over Time?",x = "Year",y = "Number of Records (Medals or Entries)",color = "Sport")+theme_minimal()

## Are Some Sports Becoming More Popular Over Time?



##which country won the most medal

data=data%>%mutate(Total=ifelse(Medal=="Gold",1,ifelse(Medal=="Silver",1,ifelse(Medal=="Bronze",1,0))))

medals_summary=data%>%group_by(Country)%>%summarize(Gold_Medals=sum(Medal=="Gold"),Silver_Medals=sum(Medal=="Silver"),Bronze_Medals=sum(Medal=="Bronze"),Total_Medals=sum(Total))

sorted_medals=medals_summary%>%arrange(desc(Total_Medals))

```
top_countries=sorted_medals%>%top_n(10,Total_Medals)

top_countries
```

# A tibble: 10 × 5

| Country | Gold_Medals | Silver_Medals | Bronze_Medals | Total_Medals |
|---|---|---|---|---|
| 1 United States | 928 | 583 | 481 | 1992 |
| 2 Soviet Union | 439 | 285 | 297 | 1021 |
| 3 Australia | 216 | 270 | 312 | 798 |
| 4 Germany | 237 | 176 | 278 | 691 |
| 5 China | 234 | 252 | 193 | 679 |
| 6 Russia | 192 | 206 | 240 | 638 |
| 7 East Germany | 286 | 190 | 150 | 626 |
| 8 Italy | 145 | 163 | 178 | 486 |
| 9 Romania | 135 | 157 | 190 | 482 |
| 10 United Kingdom | 122 | 157 | 188 | 467 |

```
ggplot(top_countries, aes(x = reorder(Country, Total_Medals), y = Total_Medals))
+geom_bar(stat = "identity", fill = "Pink") +coord_flip() +labs(title = "Top Countries by Total
Olympic Medals",x = "Country", y = "Total Medals") +theme_minimal()
```

Top Countries by Total Olympic Medals
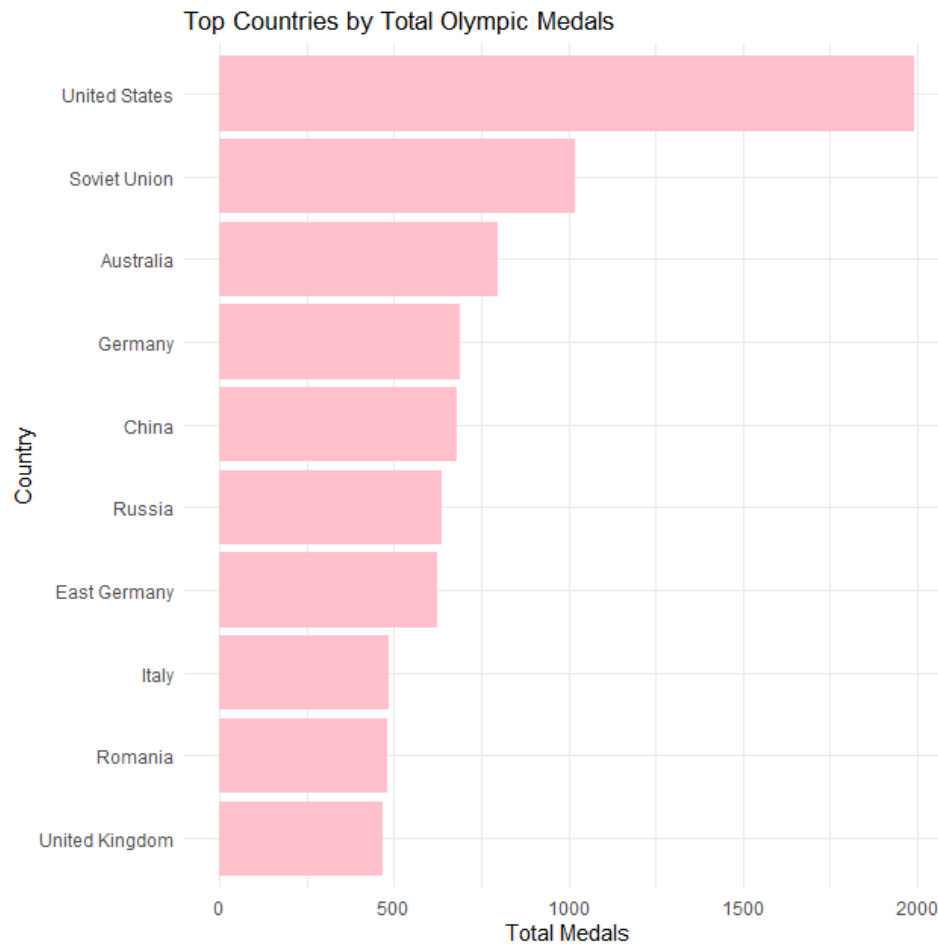
## Who are the most successful athletes?

```
top_athletes=data%>%group_by(Athlete)%>%summarise(Total_Medals=n())%>%arrange(
desc(Total_Medals))%>%slice_head(n=10) top_athletes
```

# A tibble: 10 × 2

| Athlete | Total_Medals |
| --- | --- |
| 1 PHELPS, Michael | 16 |
| 2 ANDRIANOV, Nikolay | 12 |

| | | |
|---|---|---|
| 3 FISCHER, Birgit | 12 |
| 4 NEMOV, Alexei | 12 |
| 5 THOMPSON, Jenny | 12 |
| 6 TORRES, Dara | 12 |
| 7 BIONDI, Matthew | 11 |
| 8 COUGHLIN, Natalie | 11 |
| 9 DITYATIN, Aleksandr | 10 |
| 10 HALL, Gary Jr. | 10 |

```
ggplot(top_athletes, aes(x = reorder(Athlete, Total_Medals), y = Total_Medals))
+geom_bar(stat = "Identity", color = "black",fill="pink") +coord_flip() +labs(title = "Top 10
Most Successful Olympic Athletes",x = "Athlete", y = "Total Medals") +theme_minimal()
```

## Top 10 Most Successful Olympic Athletes
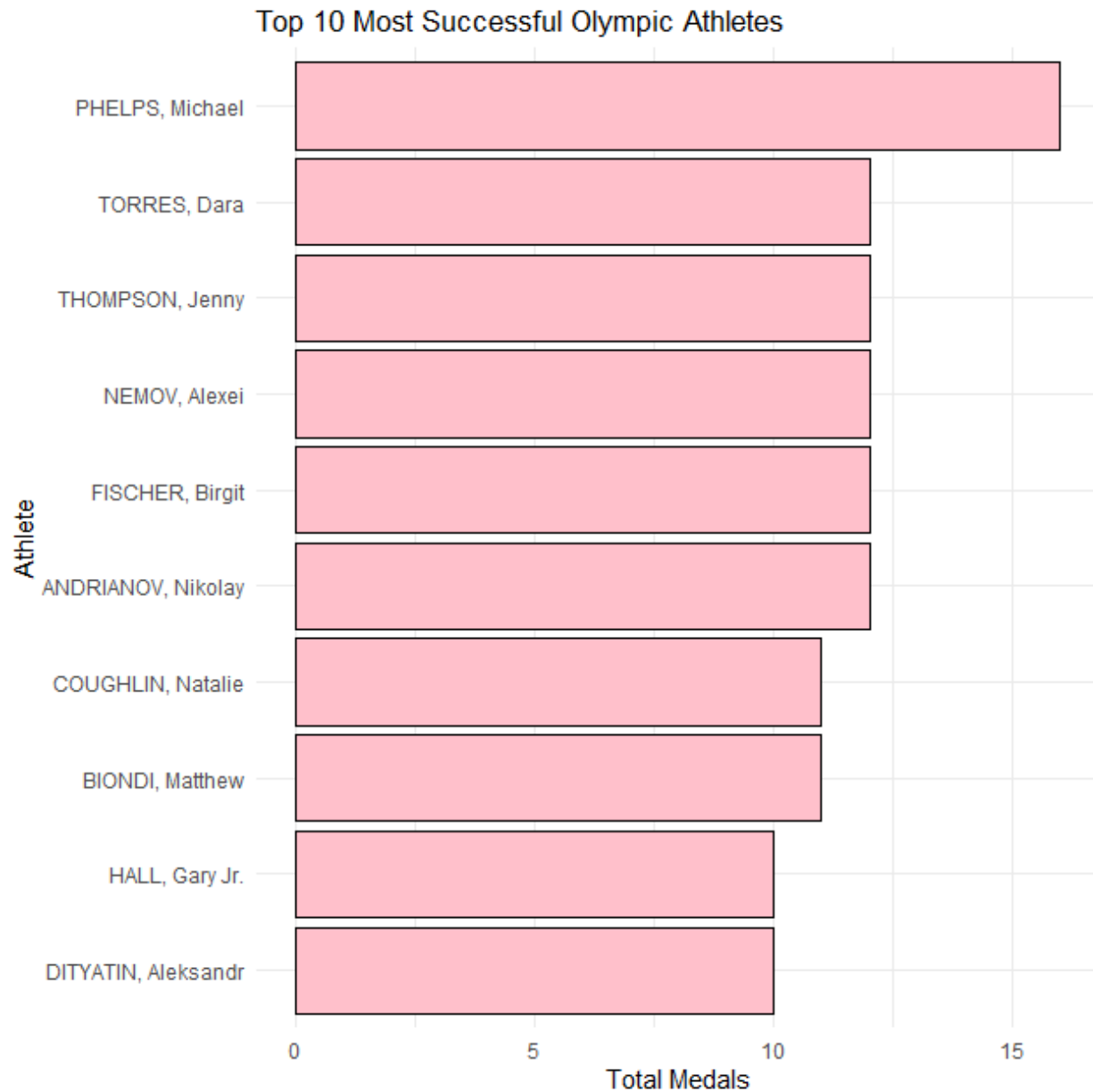


## What are the top sports for each country?

top_sports_per_country=country_sports_medals%>%group_by(Country)%>%slice_head(n=1)

top_countries=top_sports_per_country%>%group_by(Country)%>%summarise(Total=sum(Total_Medals))%>%arrange(desc(Total))%>%slice(1:20)

top20=top_sports_per_country%>%filter(Country%in%top_countries$Country)

# A tibble: 20 × 3

# Groups: Country [20]

Country Sport Total_Medals 1 Australia Aquatics 239 2 Brazil Football 109 3 Canada Rowing 99 4 China Aquatics 118 5 Cuba Baseball 111 6 East Germany Rowing 146 7 France Fencing 115 8 Germany Aquatics 100 9 Hungary Canoe / Kayak 93 10 Italy Fencing 110 11 Jamaica Athletics 88 12 Japan Aquatics 70 13 Korea, South Handball 96 14 Netherlands Hockey 159 15 Romania Rowing 156 16 Russia Aquatics 125 17 Soviet Union Athletics 115 18 United Kingdom Rowing 99 19 United States Aquatics 578 20 Yugoslavia Basketball 84

```
ggplot(top20, aes(x = Total_Medals, y = reorder(Country, Total_Medals), fill = Sport))
+geom_bar(stat="identity", position="dodge") +labs(title = "Top 20 countries by Total
Medals(Grouped by sport)",x = "Total Medals", y = "Country",fill="Sport") +
theme_minimal()+theme(axis.text.y=element_text(size=10),plot.title=element_text(face="
bold",hjust=0.5))
```

**Top 20 countries by Total Medals(Grouped by sport)**
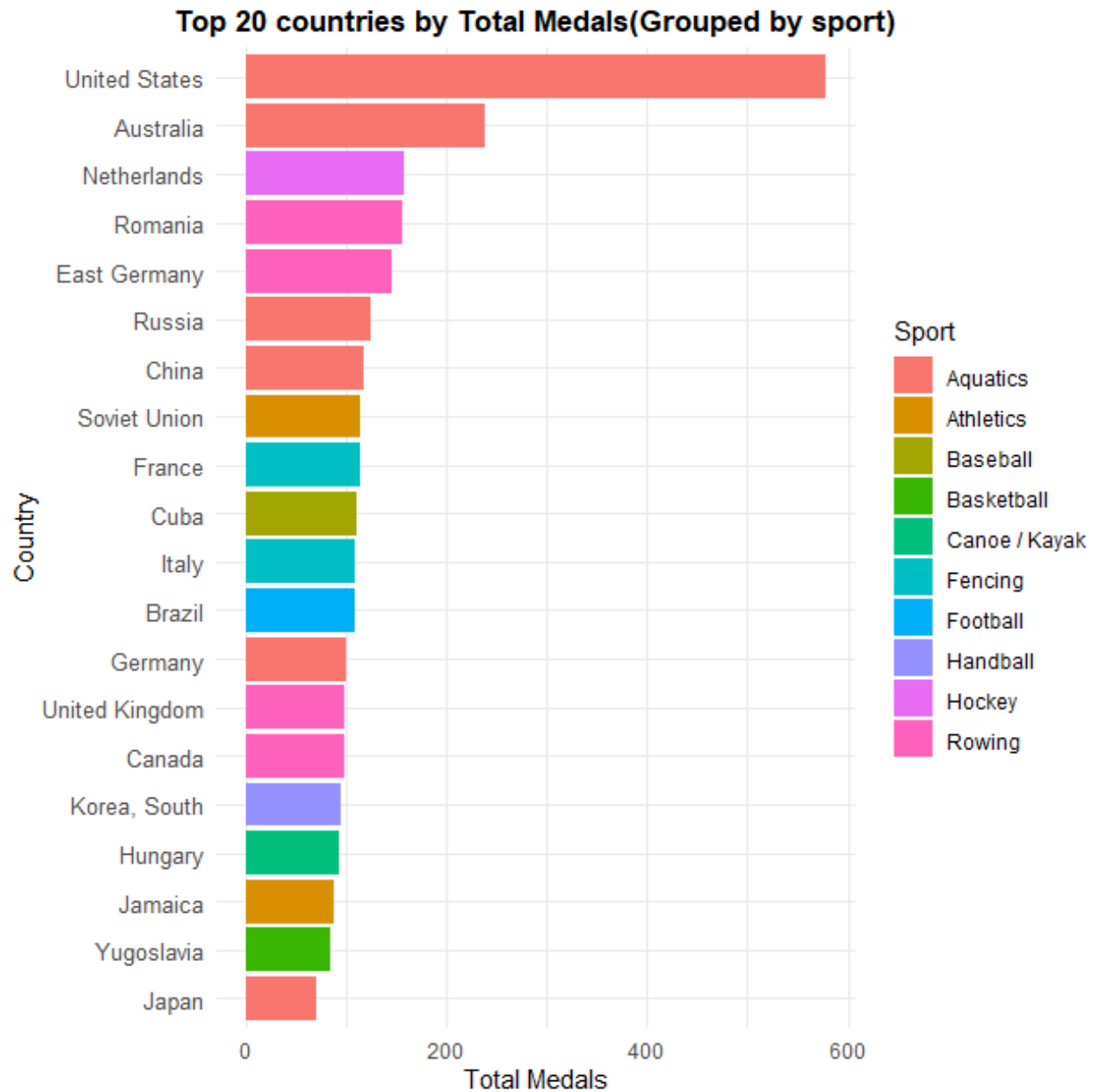
## How many male vs female athletes won medals?

gender_medals=data%>%filter(Medal%in%c("Gold","Silver","Bronze"))%>%group_by(Gender)%>%reframe(Medal=n())%>%ungroup()

gender_medals

# A tibble: 2 × 2

Gender Medal

 1 Men 9388

 2 Women 5928

ggplot(gender_medals, aes(x = "", y = Medal, fill = Gender)) +geom_bar(stat = "identity", width = 1) +coord_polar(theta = "y") +geom_text(aes(label = paste0(round(Medal / sum(Medal) * 100, 1), "%")),position = position_stack(vjust = 0.5), color = "white") +labs(title = "Proportion of Medals Won by Male vs Female Athletes") +theme_minimal() +theme(axis.text.x = element_blank(), axis.title.x = element_blank())

## Proportion of Medals Won by Male vs Female Athletes
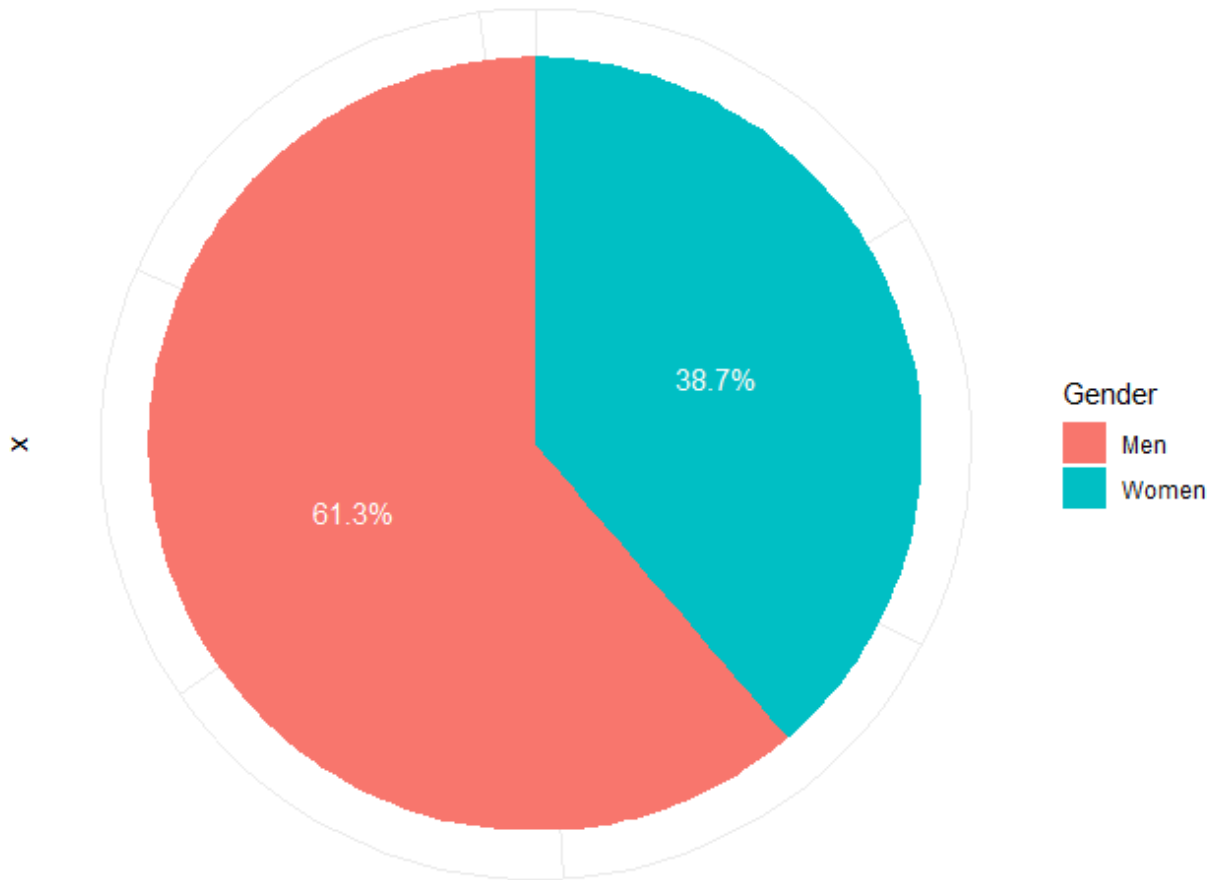


## Which sports/events are more balanced in terms of gender?

library(dplyr)

library(tidyr)

gender_balance=data%>%group_by(Sport,Gender)%>%summarise(count=n(),.groups="drop")%>%tidyr::pivot_wider(names_from=Gender,values_from=count,values_fill=0)%>%mutate(Total= Men+Women,Gender_ratio=abs(Men-Women)/Total)

```
balanced_sports=gender_balance%>%arrange(Gender_ratio)%>%head(10)
balanced_sports
```

# A tibble: 10 × 5

| Sport | Men | Women | Total | Gender_ratio |
|-------|-----|-------|-------|--------------|
| 1 Archery | 81 | 81 | 162 | 0 |
| 2 Badminton | 60 | 60 | 120 | 0 |
| 3 Basketball | 323 | 323 | 646 | 0 |
| 4 Table Tennis | 60 | 60 | 120 | 0 |
| 5 Taekwondo | 40 | 40 | 80 | 0 |
| 6 Triathlon | 9 | 9 | 18 | 0 |
| 7 Volleyball | 347 | 348 | 695 | 0.00144 |
| 8 Handball | 393 | 387 | 780 | 0.00769 |
| 9 Tennis | 60 | 58 | 118 | 0.0169 |
| 10 Gymnastics | 375 | 408 | 783 | 0.0421 |

```
library(ggplot2)

ggplot(gender_balance, aes(x = reorder(Sport, Gender_ratio), y = Gender_ratio))
+geom_bar(stat = "identity", fill = "violet") +coord_flip() +labs(title = "Gender Balance by
Sport", x = "Sport",y = "Gender Balance Ratio (Lower = More Balanced)")
```

## Gender Balance by Sport

Sport (top to bottom): Softball, Boxing, Baseball, Wrestling, Modern Pentathlon, Sailing, Weightlifting, Cycling, Canoe / Kayak, Football, Fencing, Shooting, Equestrian, Judo, Rowing, Athletics, Hockey, Aquatics, Gymnastics, Tennis, Handball, Volleyball, Triathlon, Taekwondo, Table Tennis, Basketball, Badminton, Archery

Gender Balance Ratio (Lower = More Balanced)