Arshdeep Kaur

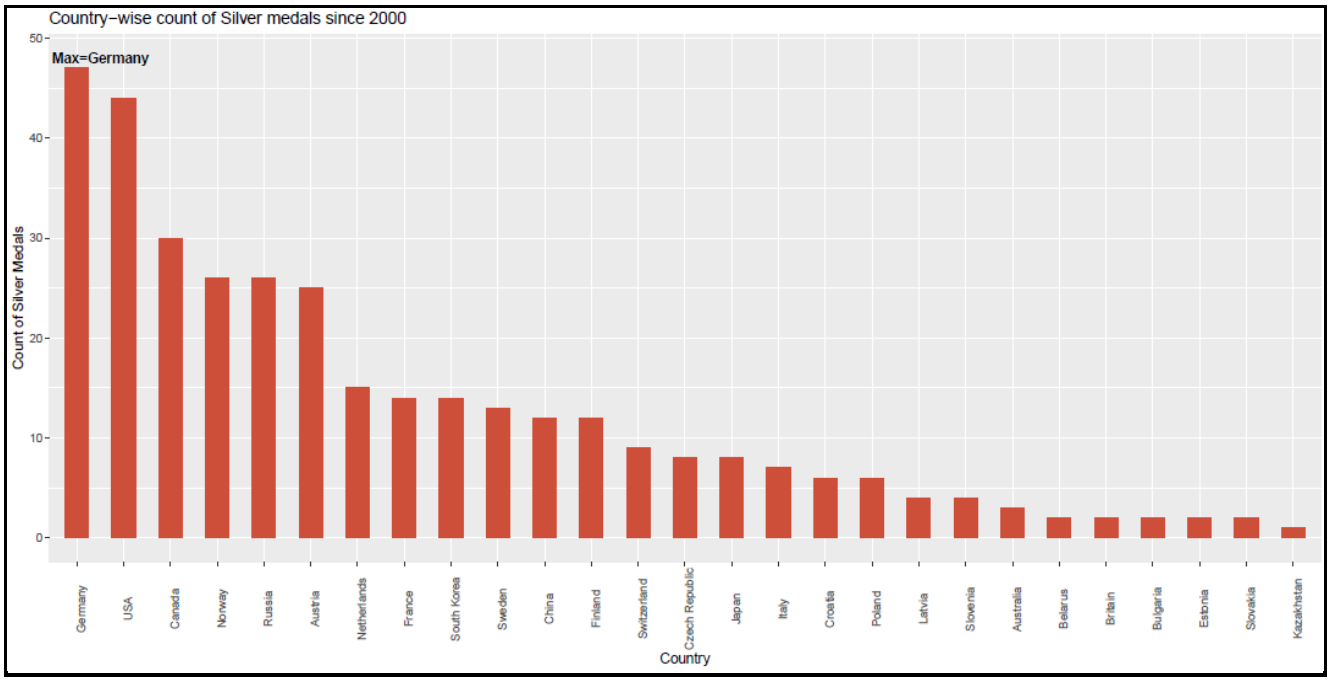
Missing Values:

Screenshot of Olympics Dataset after Imputing Missing 'ageofathlete' column:

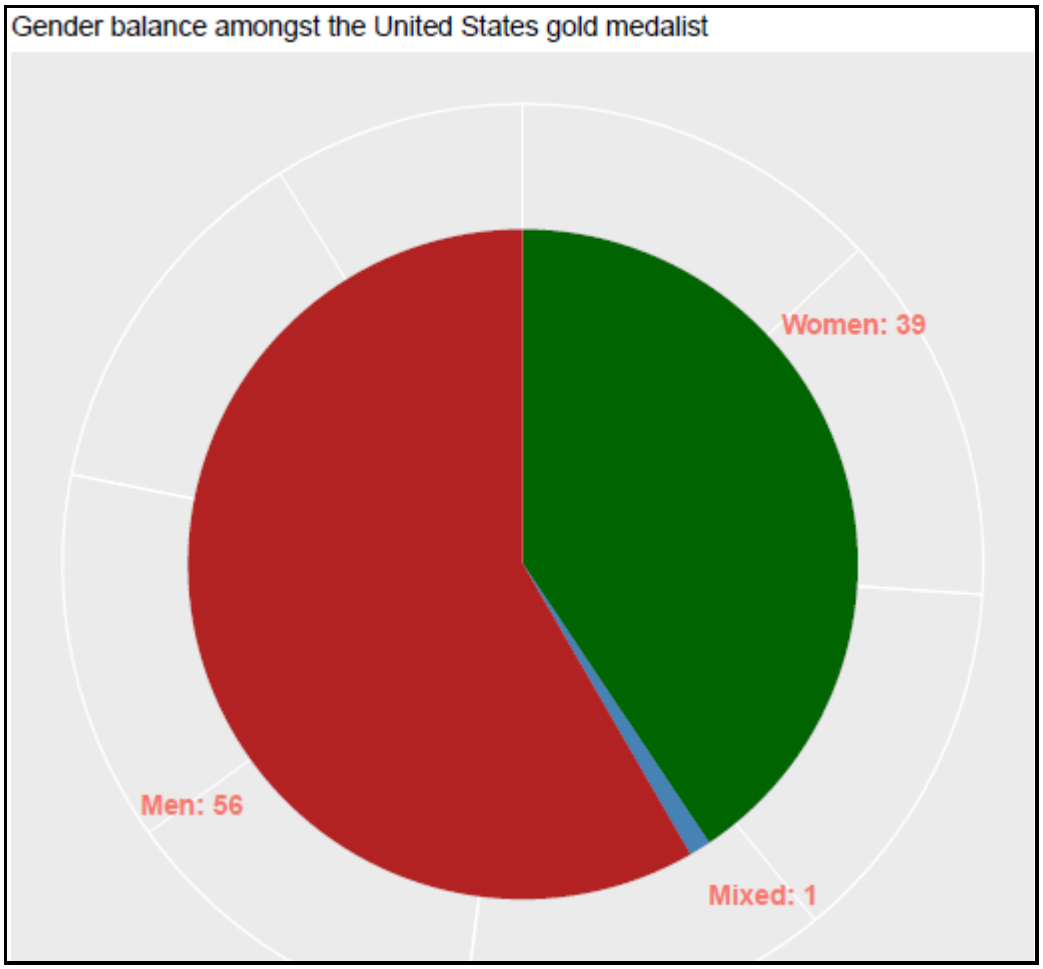
	year	sport	event	country	gender	medal_rank	medal	nameofathleteorteam	ageofathlete
1	1924	Bobsled	Men's Four/Five	Switzerland	Men	1	gold	Switzerland-1	25.09
2	1924	Bobsled	Men's Four/Five	Britain	Men	2	silver	Britain-1	26.00
3	1924	Bobsled	Men's Four/Five	Belgium	Men	3	bronze	Belgium-1	24.00
4	1924	Cross-Country Skiing	Men's 18 Kilometers	Norway	Men	1	gold	Thorleif Haug	29.00
5	1924	Cross-Country Skiing	Men's 18 Kilometers	Norway	Men	2	silver	Johan GrÃ¸ttumsbraaten	24.00
6	1924	Cross-Country Skiing	Men's 18 Kilometers	Finland	Men	3	bronze	Tapani Niku	28.00
7	1924	Cross-Country Skiing	Men's 50 Kilometers	Norway	Men	1	gold	Thorleif Haug	29.00
8	1924	Cross-Country Skiing	Men's 50 Kilometers	Norway	Men	2	silver	Thoralf StrÃ¸mstad	27.00
9	1924	Cross-Country Skiing	Men's 50 Kilometers	Norway	Men	3	bronze	Johan GrÃ¸ttumsbraaten	24.00
10	1924	Curling	Men's Curling	Britain	Men	1	gold	Britain	26.00
11	1924	Curling	Men's Curling	Sweden	Men	2	silver	Sweden	26.70
12	1924	Curling	Men's Curling	France	Men	3	bronze	France	24.00
13	1924	Figure Skating	Men's Singles	Sweden	Men	1	gold	Gillis GrafstrÃ¶m	30.00
14	1924	Figure Skating	Men's Singles	Austria	Men	2	silver	Willy BÃ¶ckl	30.00
15	1924	Figure Skating	Men's Singles	Switzerland	Men	3	bronze	Georges Gautschi	19.00
16	1924	Figure Skating	Mixed Pairs	Austria	Mixed	1	gold	Austria	24.84
17	1924	Figure Skating	Mixed Pairs	Finland	Mixed	2	silver	Finland	26.59
18	1924	Figure Skating	Mixed Pairs	France	Mixed	3	bronze	France-1	24.00

Showing 1 to 18 of 2,865 entries

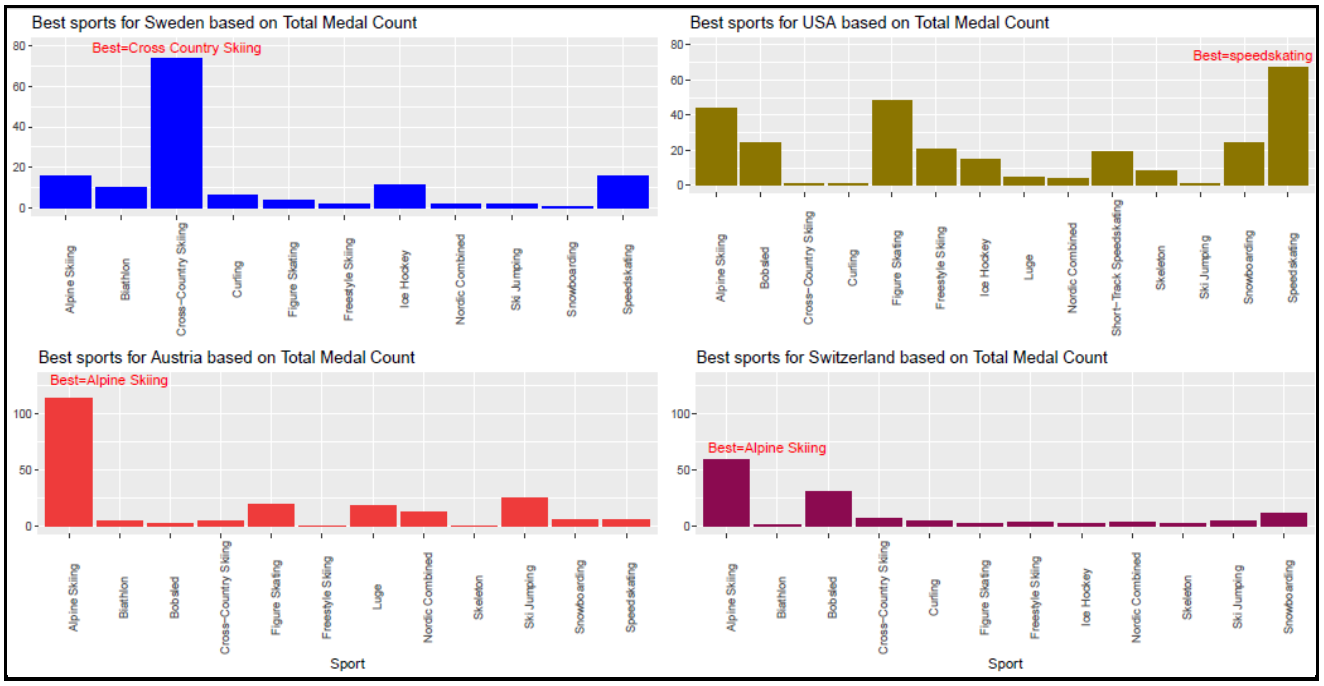
What country has won most Silver medals since 2000?



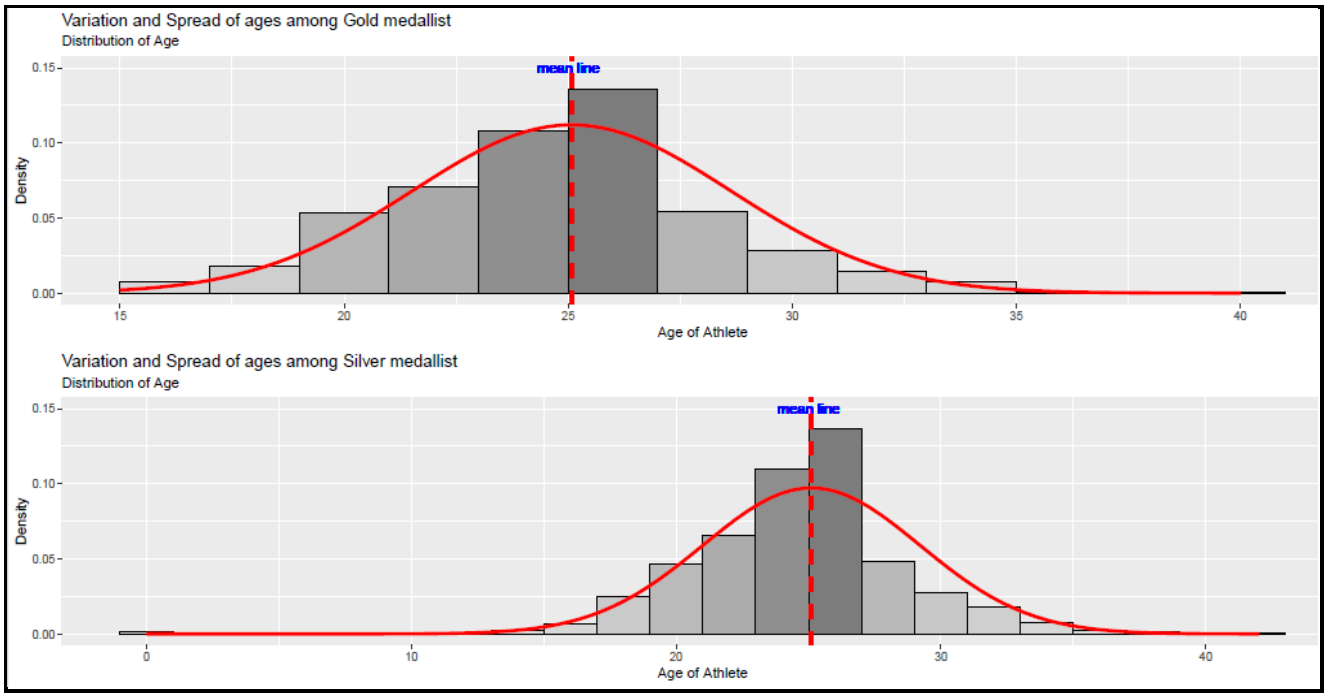
How is the gender balance amongst the United States gold medalist?



What are the best sports for Sweden, USA, Austria and Switzerland?



What is the variation and spread of ages amongst gold and silver medalists?



Appendices-Code

```
# load library for plotting the data

library('ggplot2')
```

```
# load library to perform sql function on dataframe
library('sqldf')
# import the dataset
olympicdata<-read.csv('OlympicGames.csv')
# view the dataframe containing imported data
View(olympicdata)
```

 # Impute Missing Values

```
mean_data<-aggregate(data=olympicdata,Age.of.Athlete~Country,FUN = mean)
mean_data$Age.of.Athlete<-round(mean_data$Age.of.Athlete,2)
View(mean_data)


for(i in 1:nrow(mean_data)){
for(j in 1:nrow(olympicdata)){
  if((olympicdata$Country[j]==mean_data$Country[i]) & (is.na(olympicdata$Age.of.Athlete[j]))){
    {
      olympicdata$Age.of.Athlete[j]<-mean_data$Age.of.Athlete[i]
    }
  }
 }
}


for(j in 1:nrow(olympicdata)){
  if( (is.na(olympicdata$Age.of.Athlete[j])) ){
    {
      olympicdata$Age.of.Athlete[j]<-0
    }
  }
}
View(olympicdata)
```

Question 1: What country has won most Silver medals since 2000?

```
most_silver<-sqldf("select Country, count(Medal) as count from olympicdata where Medal=='silver' and Year>=2000 group by Country")

# sort dataframe in descending order

most_silver <- most_silver[with(most_silver, order(-count)), ]

View(most_silver)

#plot the data

ggplot(data = most_silver)+
  labs(x='Country',y='Count of Silver Medals',title='Country-wise count of Silver medals since 2000')+
  geom_bar(stat = 'identity',aes(x=most_silver$Country,y=most_silver$count),width=.5,fill='tomato3') +
  annotate("text", colour='black',label='Max=Germany', x = 1.5, y = 48, fontface="bold")+
  theme(axis.text.x = element_text(angle = 90))+
  scale_x_discrete(limits= most_silver$Country)
```

Question 2: How is the gender balance amongst the United States gold medalist?

```
usgold_data<-sqldf("select Gender from olympicdata where Country='USA' and Medal='gold' ")

usgold_data<-as.data.frame(table(usgold_data))

colnames(usgold_data)<-c('gender','count')

View(usgold_data)


# create pie-chart for the data


# val<-sort(usgold_data$count,decreasing = TRUE)

# val<-rev(usgold_data$count)

# typeof(val)

val<-c('Women: 39','Mixed: 1','Men: 56')

ggplot(data = usgold_data)+
  labs(x=NULL,y=NULL, title='Gender balance amongst the United States gold medalist')+
  geom_bar(stat = 'identity',aes(x="",y=count,fill = factor(gender)))+ #,levels = rev(as.character(gender)))))+
  coord_polar(theta = "y",start = 0)+
  scale_fill_manual(values = c("Men"='firebrick',"Mixed"='steelblue',"Women"='darkgreen'))+
  guides(fill = guide_legend(title = "Gender"))+
```

```
 geom_text(aes(x=1.65,y=24,label=paste(val),size=20,colour='white',fontface = "bold"),position = position_stack(vjust = 0.6))+
 theme(axis.ticks = element_blank(),
      axis.text = element_blank(),
      axis.title = element_blank(),
      legend.position = "none")
```

Question 3: What are the best sports for Sweden, USA, Austria and Switzerland?

```
df1<-sqldf("select Sport,count(Medal) from olympicdata where Country='USA' group by Sport")
df2<-sqldf("select Sport,count(Medal) from olympicdata where Country='Sweden' group by Sport")
df3<-sqldf("select Sport,count(Medal) from olympicdata where Country='Austria' group by Sport")
df4<-sqldf("select Sport,count(Medal) from olympicdata where Country='Switzerland' group by Sport")
colnames(df1)<-c('sport','count')
colnames(df2)<-c('sport','count')
colnames(df3)<-c('sport','count')
colnames(df4)<-c('sport','count')

p1<-ggplot(data = df2,aes(x=sport,y=count))
p1<-p1+
 labs(title='Best sports for Sweden based on Total Medal Count')+ylim(0,80)+
 geom_bar(stat = 'identity',fill='blue')+
 theme(axis.title.y = element_blank(),axis.title.x = element_blank(),axis.text.x = element_text(angle = 90))+
 annotate("text", colour='red',label='Best=Cross Country Skiing', x = 3, y = 79)

p2<-ggplot(data = df1,aes(x=sport,y=count))
p2<-p2+labs(title='Best sports for USA based on Total Medal Count')+ylim(0,80)+
 geom_bar(stat = 'identity',fill='gold4')+
 theme(axis.title.y = element_blank(),axis.title.x = element_blank(),axis.text.x = element_text(angle = 90))+
 annotate("text", colour='red',label='Best=speedskating', x = 13.2, y = 74)

p3<-ggplot(data = df3,aes(x=sport,y=count))
p3<-p3+labs(x='Sport', title='Best sports for Austria based on Total Medal Count')+ylim(0,130)+
```

```
  geom_bar(stat = 'identity',fill='brown2')+
  theme(axis.text.x = element_text(angle = 90),axis.title.y = element_blank())+
  annotate("text", colour='red',label='Best=Alpine Skiing', x = 1.8, y = 130)


p4<-ggplot(data = df4,aes(x=sport,y=count))
p4<-p4+labs(x='Sport', title='Best sports for Switzerland based on Total Medal Count')+ylim(0,130)+
  geom_bar(stat = 'identity',fill='deeppink4')+
  theme(axis.text.x = element_text(angle = 90),axis.title.y = element_blank())+
  annotate("text", colour='red',label='Best=Alpine Skiing', x = 1.8, y = 70)


grid.arrange(p1,p2,p3,p4,nrow=2,ncol=2)
```

Question 4: What is the variation and spread of ages amongst gold and silver medalists?

```
colnames(olympicdata)<-
c('year','sport','event','country','gender','medal_rank','medal','nameofathleteorteam','ageofathlete')
age_data<-sqldf("select ageofathlete,medal from olympicdata where medal='gold' or medal='silver' order by medal")
View(age_data_silver)
age_data_gold<-subset(age_data,age_data$medal=='gold')
age_data_silver<-subset(age_data,age_data$medal=='silver')


# Create the histogram to check for normal disribution


gg<-ggplot(age_data_gold,aes(x=ageofathlete))
# change label for x-axis and y-axis
gg<-gg+labs(x='Age of Athlete',y='Density')
# manage binwidth and colours for histogram
gg<-gg+geom_histogram(binwidth = 2,colour='black',aes(y=..density..,fill=..count..))
gg<-gg+scale_fill_gradient("Count",low="#DCDCDC", high="#7C7C7C")
#adding normal curve to histogram
gg<-gg+stat_function(fun = dnorm,color="red",size=1.2,args =
list(mean=mean(age_data_gold$ageofathlete,na.rm = TRUE),sd=sd(age_data_gold$ageofathlete,na.rm =
TRUE)))
```

```
# add title to the plot

gg<-gg+ggtitle('Variation and Spread of ages among Gold medallist',subtitle = 'Distribution of Age')

gg<-
gg+geom_vline(data=age_data_gold,xintercept=mean(age_data_gold$ageofathlete),linetype='dashed',size=1.
2,color='red')+

   geom_text(aes(x=25, label="mean line", y=.15), colour="blue", angle=0, text=element_text(size=1.2))+

 theme(legend.position="none")


gg1<-ggplot(age_data_silver,aes(x=ageofathlete))

# change label for x-axis and y-axis

gg1<-gg1+labs(x='Age of Athlete',y='Density')

# manage binwidth and colours for histogram

gg1<-gg1+geom_histogram(binwidth = 2,colour='black',aes(y=..density..,fill=..count..))

gg1<-gg1+scale_fill_gradient("Count",low="#DCDCDC", high="#7C7C7C")

#adding normal curve to histogram

gg1<-gg1+stat_function(fun = dnorm,color="red",size=1.2,args =
list(mean=mean(age_data_silver$ageofathlete,na.rm = TRUE),sd=sd(age_data_silver$ageofathlete,na.rm =
TRUE)))

# add title to the plot

gg1<-gg1+ggtitle('Variation and Spread of ages among Silver medallist',subtitle = 'Distribution of Age')+


geom_vline(data=age_data_silver,xintercept=mean(age_data_silver$ageofathlete),linetype='dashed',size=1.2,
color='red')+

 geom_text(aes(x=25, label="mean line", y=.15), colour="blue", angle=0, text=element_text(size=1.2)) +

 theme(legend.position="none")


grid.arrange(gg,gg1,nrow=2,ncol=1)
```