STAT 5730 final project

James Laughead

4/17/2022

```
#Loading in the data
library(readr)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5     v dplyr   1.0.7
## v tibble  3.1.5     v stringr 1.4.0
## v tidyr   1.1.4     v forcats 0.5.1
## v purrr   0.3.4
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
athlete_events <- read_csv(
   file = 'athlete_events.csv',
   col_types = cols(ID = 'i', Age = 'i', Height = 'i', Year = 'i')
)
noc_regions <- read_csv('noc_regions.csv')
```

```
## Rows: 230 Columns: 3
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (3): NOC, region, notes
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
athlete_events %>%
  filter(!is.na(Medal)) %>%
  count(Games, Event, NOC, Medal)
```

```
## # A tibble: 18,905 x 5
##    Games       Event                               NOC   Medal      n
##    <chr>       <chr>                               <chr> <chr>  <int>
##  1 1896 Summer Athletics Men's 1,500 metres        AUS   Gold       1
##  2 1896 Summer Athletics Men's 1,500 metres        FRA   Bronze     1
##  3 1896 Summer Athletics Men's 1,500 metres        USA   Silver     1
##  4 1896 Summer Athletics Men's 100 metres          GER   Silver     1
##  5 1896 Summer Athletics Men's 100 metres          HUN   Bronze     1
##  6 1896 Summer Athletics Men's 100 metres          USA   Bronze     1
##  7 1896 Summer Athletics Men's 100 metres          USA   Gold       1
##  8 1896 Summer Athletics Men's 110 metres Hurdles GBR   Silver     1
##  9 1896 Summer Athletics Men's 110 metres Hurdles USA   Gold       1
```

```
## 10 1896 Summer Athletics Men's 400 metres         GBR   Bronze     1
## # ... with 18,895 more rows
```

```
library(gapminder)
library(countrycode)
```

```
## Warning: package 'countrycode' was built under R version 4.1.2
```

```
host_cities <- read.csv("host_cities.csv")
#countrycode(athlete_events)
athlete_events$NOC <- countrycode(athlete_events$NOC, origin = "ioc", destination = "iso3c")
```

```
## Warning in countrycode_convert(sourcevar = sourcevar, origin = origin, destination = dest, : Some val
```

```
gapminder <- gapminder

#First, I filter out the olympic dataset so that we only have a range of values listed in the gapminder
#From module 4
athlete.events.change  <- filter(athlete_events, Year >= 1992 & Year <= 2007)

#Next, I delete odd years since olympics are not run in odd years
gapminder.even <- filter(gapminder, year %% 2 == 0)
#Then, I take out even years the olympics were not held, so that now we can do a comparison of GDP of c
gapminder.new <- filter(gapminder.even, year != 1962 & year != 1982)

#Final filter so now we only have data in which we have olympics and GDP data for
athlete.events.change1  <- filter(athlete_events, Year == 1952 | Year == 1972 | Year == 1992 | Year == 1
#Then I change the name of the dataset so I can combine them with left join
athlete.events.change1$year <- athlete.events.change1$Year

new.df <- left_join(athlete.events.change1, gapminder.new)
```

```
## Joining, by = "year"
```

```
#After this, I filter out values only such that Team == country, so we only have correct and unique val
#from module 7
new.df1 <- filter(new.df, Team == country)

#Now we have a data frame in which we can start working on the first problem with
```

```
#Is there a relationship between GDP and number of athletes?
#filter out each year individually
#Note: In some cases, there are individuals who can compete in more than one event. For the sake of thi

#1952
#Each olympics is unique, so I separated by year in order to see if there were differences among indivi
#from module 4
athletes.1952 <- filter(new.df1, Year == 1952)


gdpvsathletes1952 <- data.frame(table(athletes.1952$country))
gdpvsathletes1952$country <- gdpvsathletes1952$Var1

#created one dataframe where gdpPerCapital is included with country
#from module 7
new.1952 <- left_join(athletes.1952, gdpvsathletes1952)
```

```
## Joining, by = "country"
```

```
#I repeat the same steps for all four years
#1972
athletes.1972 <- filter(new.df1, Year == 1972)


gdpvsathletes1972 <- data.frame(table(athletes.1972$country))
gdpvsathletes1972$country <- gdpvsathletes1972$Var1

new.1972 <- left_join(athletes.1972, gdpvsathletes1972)
```

```
## Joining, by = "country"
```

```
new.1972 <- filter(new.1972, gdpPercap < 100000)


#1992
athletes.1992 <- filter(new.df1, Year == 1992)

gdpvsathletes1992 <- data.frame(table(athletes.1992$country))
gdpvsathletes1992$country <- gdpvsathletes1992$Var1

new.1992 <- left_join(athletes.1992, gdpvsathletes1992)
```

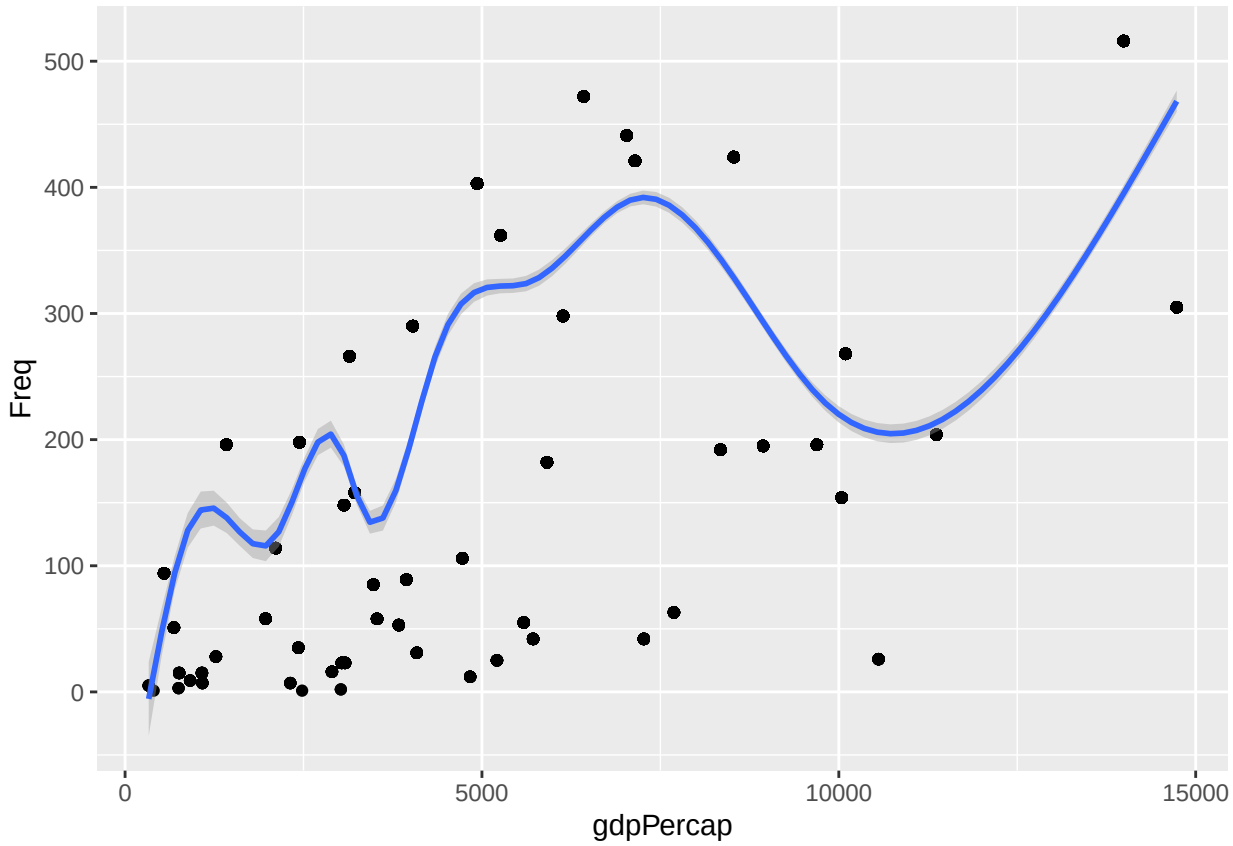
```
## Joining, by = "country"
```

```
#2002
athletes.2002 <- filter(new.df1, Year == 2002)

gdpvsathletes2002 <- data.frame(table(athletes.2002$country))
gdpvsathletes2002$country <- gdpvsathletes2002$Var1

new.2002 <- left_join(athletes.2002, gdpvsathletes2002)
```

```
## Joining, by = "country"
```

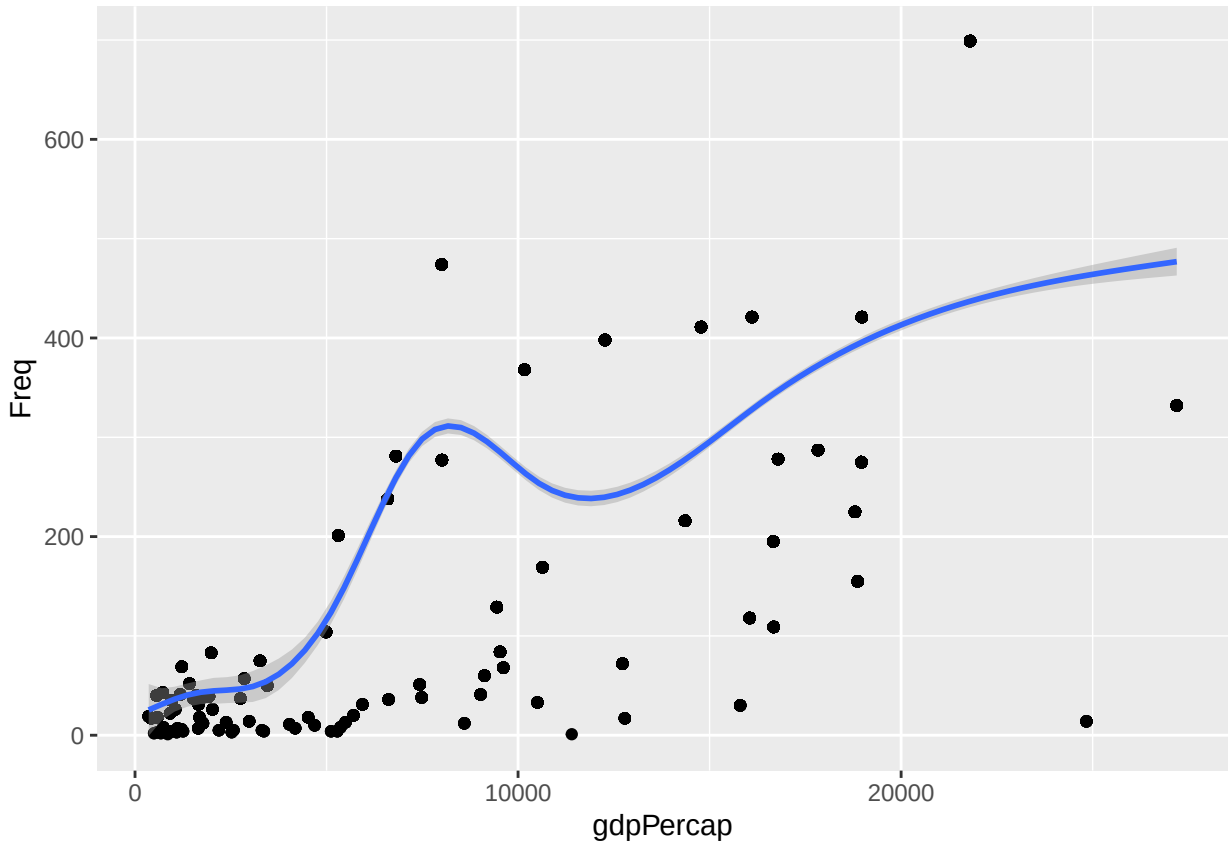
```
#Now we have each year individually, and we can start to answer our question
#I decided to plot the data, and use a smoothing line in order to make conclusions.
#from module 2
ggplot(data = new.1952) +
  geom_point(mapping = aes(x = gdpPercap, y = Freq))+
  geom_smooth(mapping = aes(x = gdpPercap, y = Freq))
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



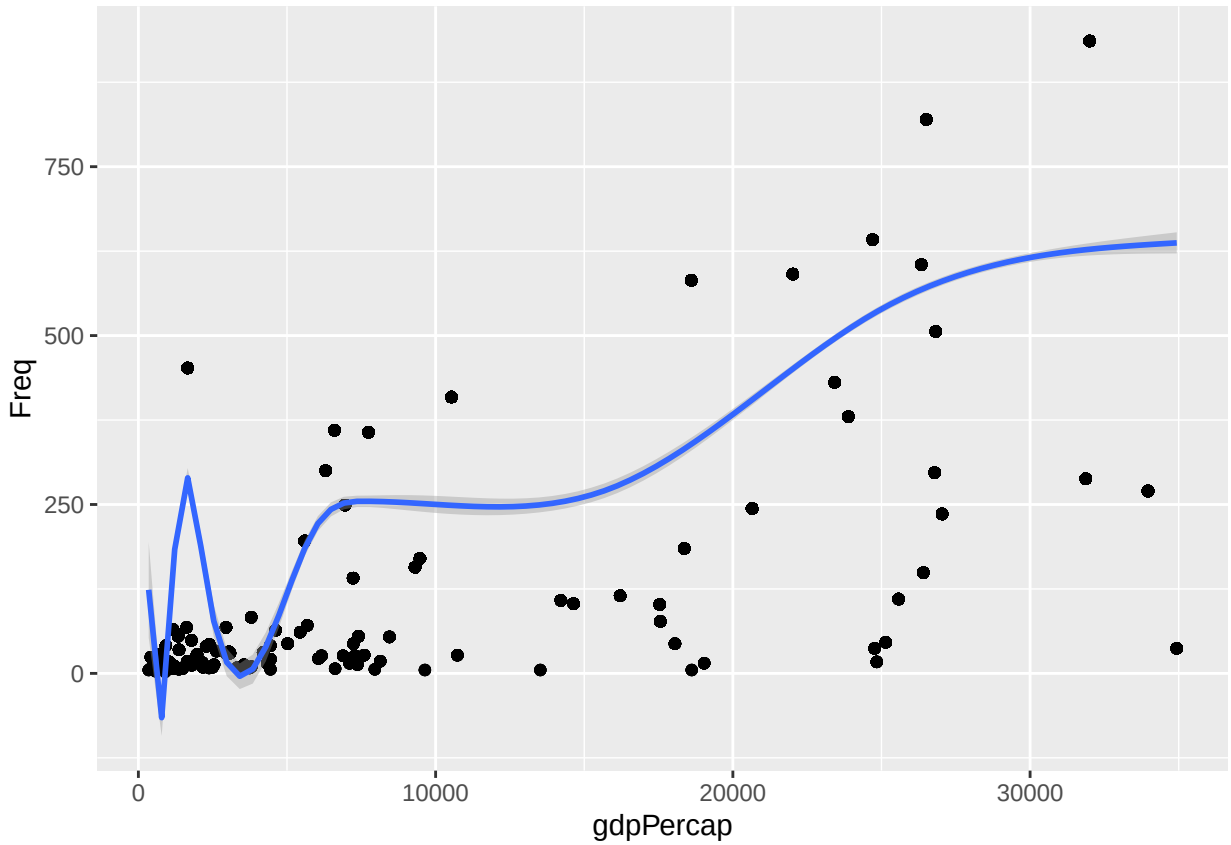

```
ggplot(data = new.1972) +
  geom_point(mapping = aes(x = gdpPercap, y = Freq))+
  geom_smooth(mapping = aes(x = gdpPercap, y = Freq))
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



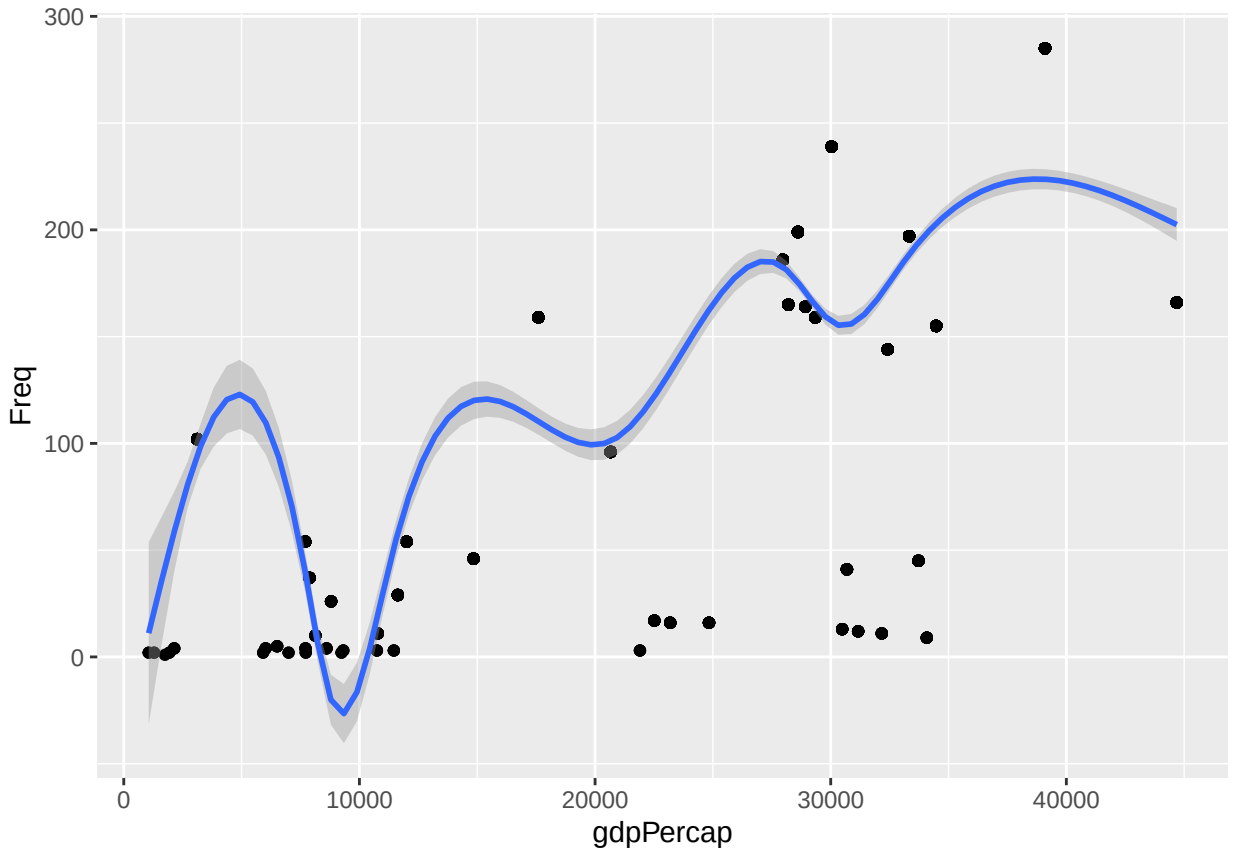

```
ggplot(data = new.1992) +
  geom_point(mapping = aes(x = gdpPercap, y = Freq))+
  geom_smooth(mapping = aes(x = gdpPercap, y = Freq))
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```




```
ggplot(data = new.2002) +
  geom_point(mapping = aes(x = gdpPercap, y = Freq))+
  geom_smooth(mapping = aes(x = gdpPercap, y = Freq))
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



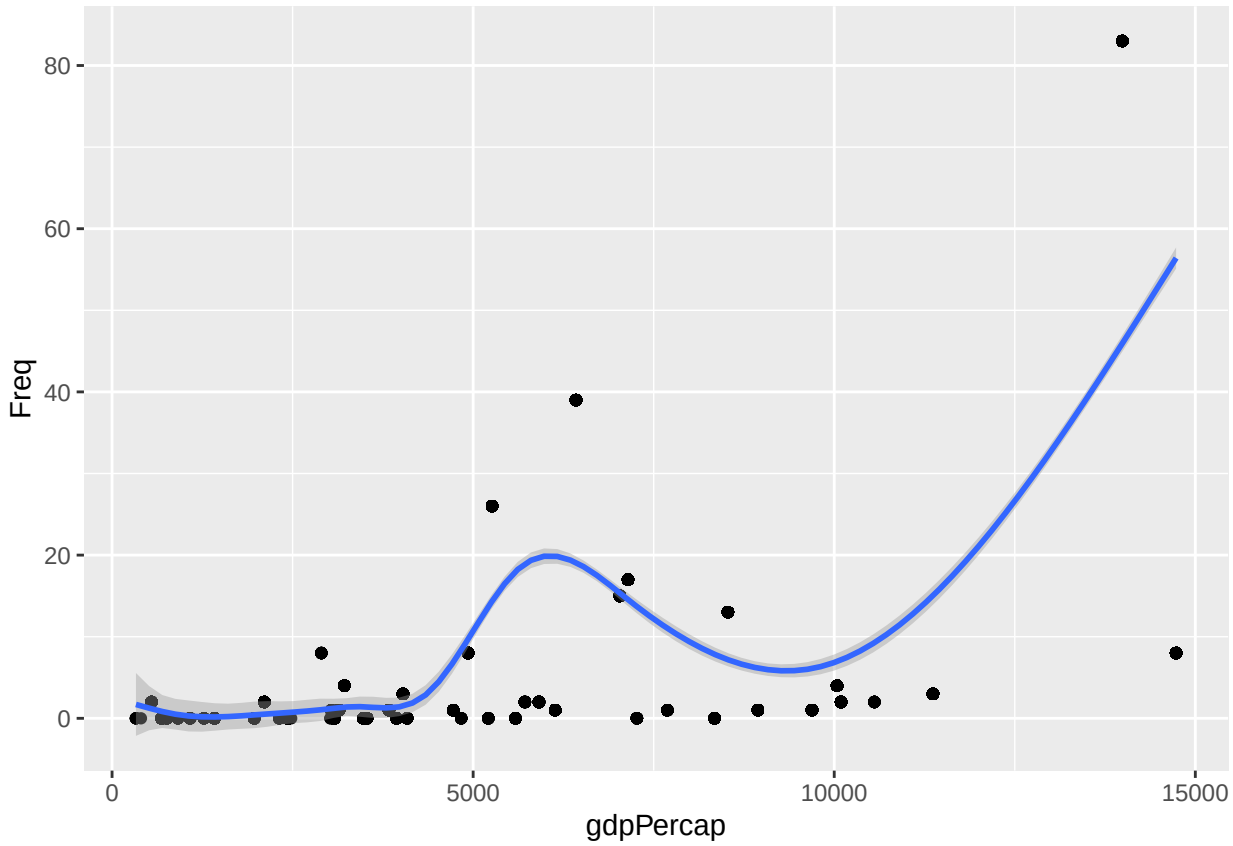

```
#In each of the graphs, there is a positive trend, which shows that countries with more gdp per capita
```

```
#Is there a relationship between GDP and number of medals?
#1952 medals by country
#Here, I use similar methods as in problem 1 in order to obtain the total medal amount for the countrie
#from module 2,4,7
medals.1952 <- na.omit(new.1952)
medals.1952 <- data.frame(table(medals.1952$country))
medals.1952$country <- medals.1952$Var1

new.medals.1952 <- left_join(athletes.1952, medals.1952)
```

```
## Joining, by = "country"
```

```
ggplot(data = new.medals.1952) +
  geom_point(mapping = aes(x = gdpPercap, y = Freq)) +
  geom_smooth(mapping = aes(x = gdpPercap, y = Freq))
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```

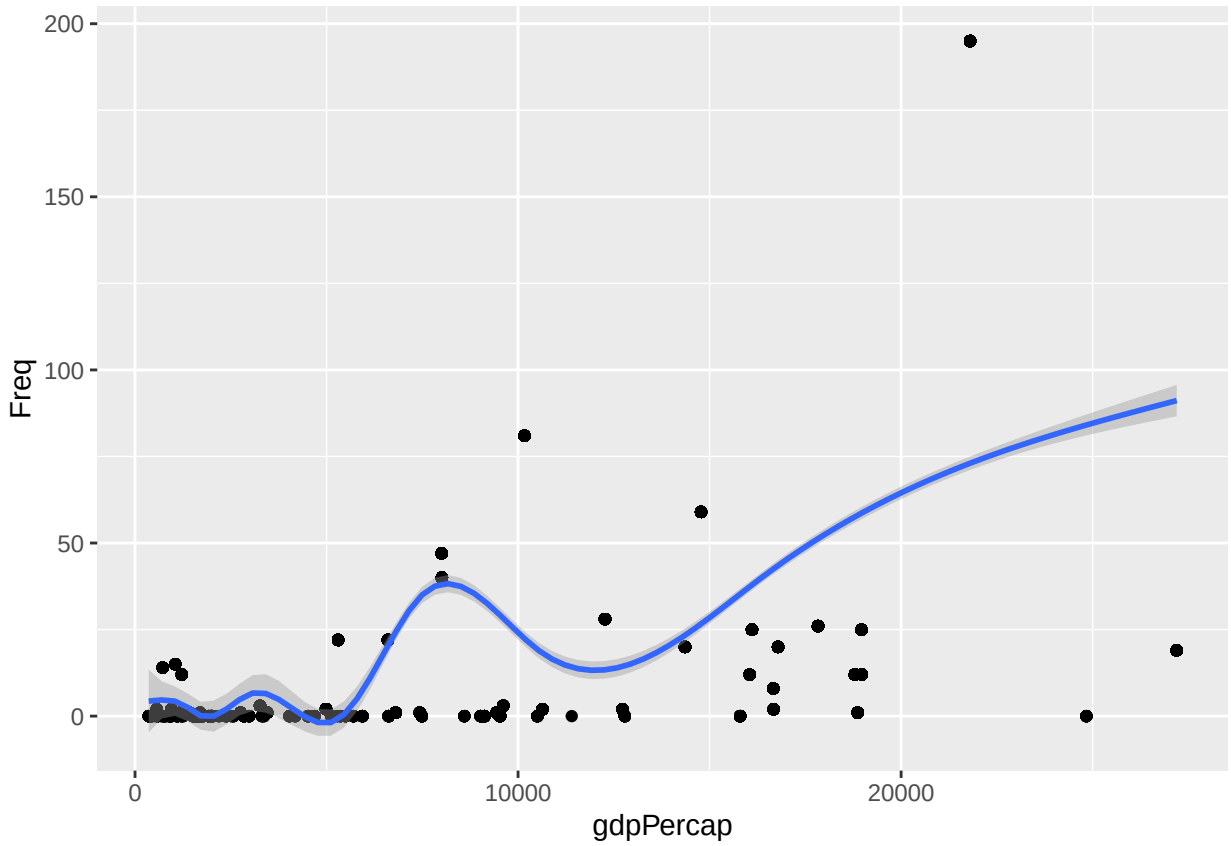



```
#Similarly to in problem 1, we plot the gdpPercap this time against the medals that country won, and we

#1972 medals by country
medals.1972 <- na.omit(new.1972)
medals.1972 <- data.frame(table(medals.1972$country))
medals.1972$country <- medals.1972$Var1

new.medals.1972 <- left_join(athletes.1972, medals.1972)
```

```
## Joining, by = "country"
```

```
new.medals.1972 <- filter(new.medals.1972, gdpPercap < 100000)

ggplot(data = new.medals.1972) +
  geom_point(mapping = aes(x = gdpPercap, y = Freq))+
  geom_smooth(mapping = aes(x = gdpPercap, y = Freq))
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```

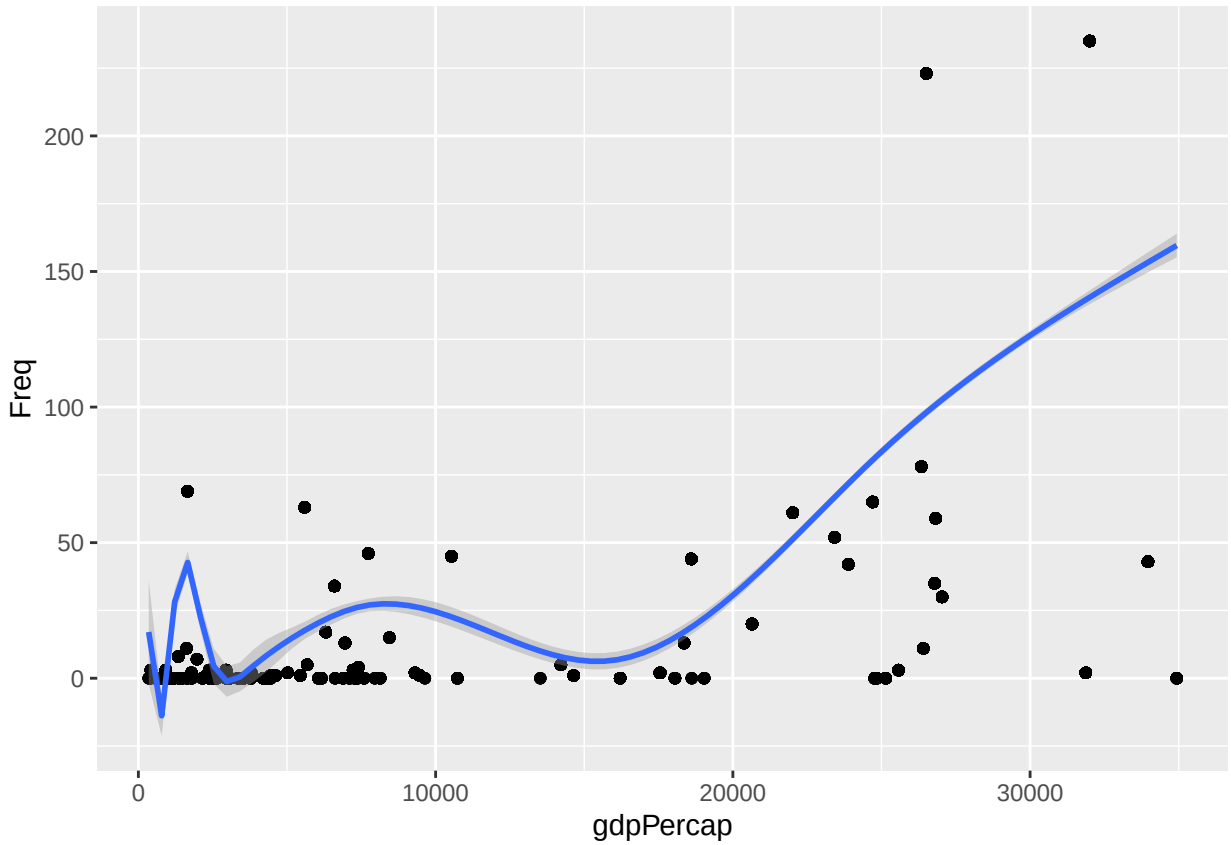



```
#1992 medals by country
medals.1992 <- na.omit(new.1992)
medals.1992 <- data.frame(table(medals.1992$country))
medals.1992$country <- medals.1992$Var1

new.medals.1992 <- left_join(athletes.1992, medals.1992)
```

```
## Joining, by = "country"
```

```
ggplot(data = new.medals.1992) +
  geom_point(mapping = aes(x = gdpPercap, y = Freq)) +
  geom_smooth(mapping = aes(x = gdpPercap, y = Freq))
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```

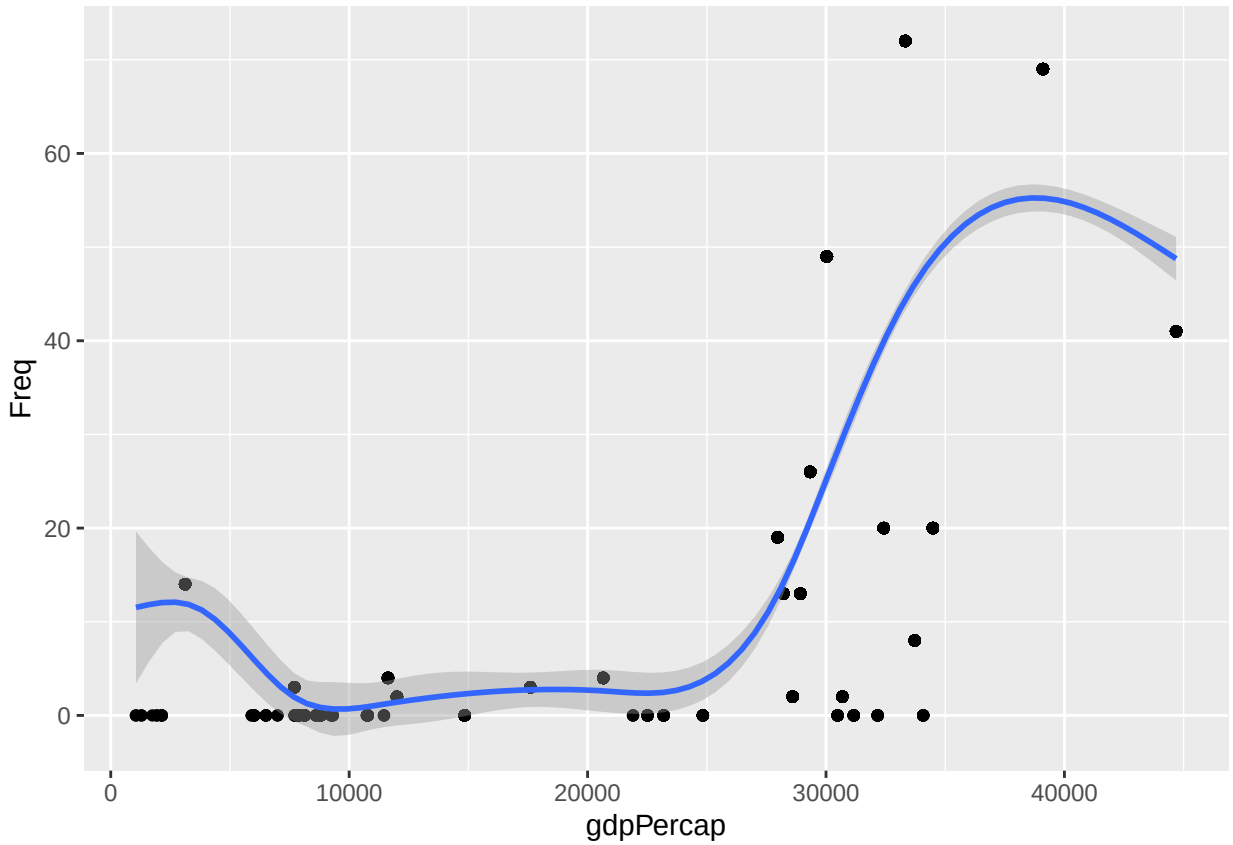



```
#2002 medals by country
medals.2002 <- na.omit(new.2002)
medals.2002 <- data.frame(table(medals.2002$country))
medals.2002$country <- medals.2002$Var1

new.medals.2002 <- left_join(athletes.2002, medals.2002)
```

```
## Joining, by = "country"
```

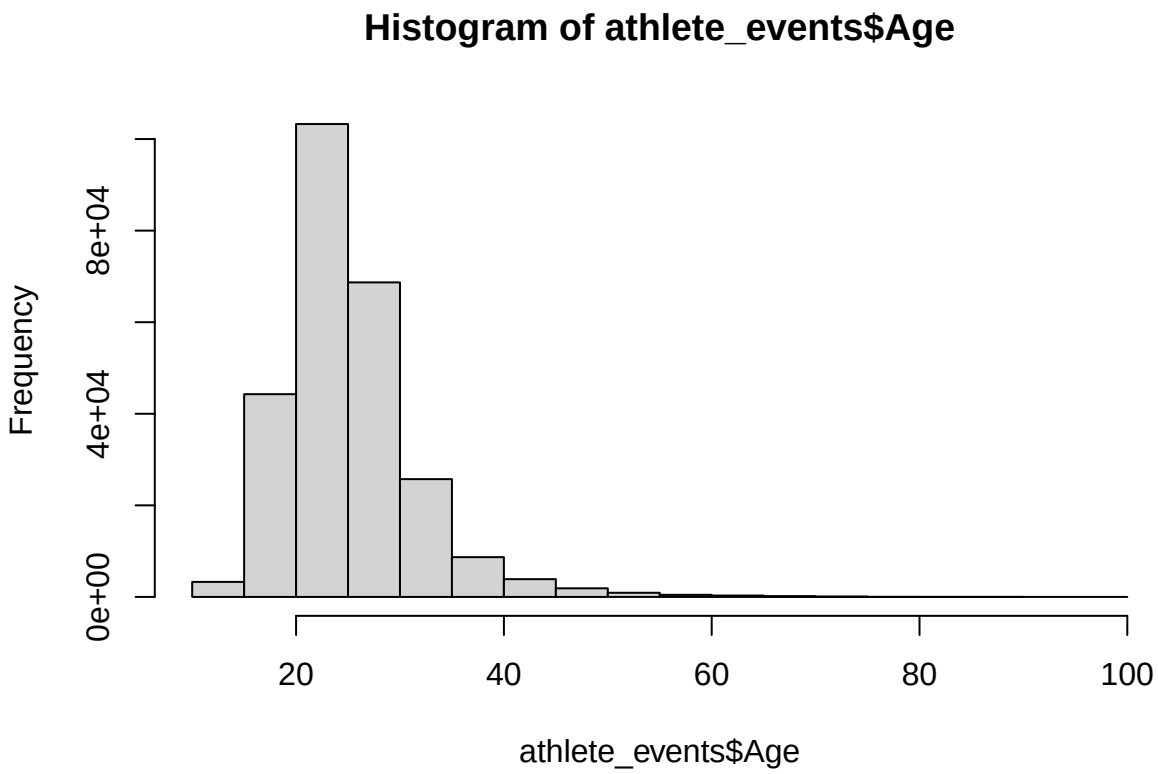
```
ggplot(data = new.medals.2002) +
  geom_point(mapping = aes(x = gdpPercap, y = Freq)) +
  geom_smooth(mapping = aes(x = gdpPercap, y = Freq))
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```

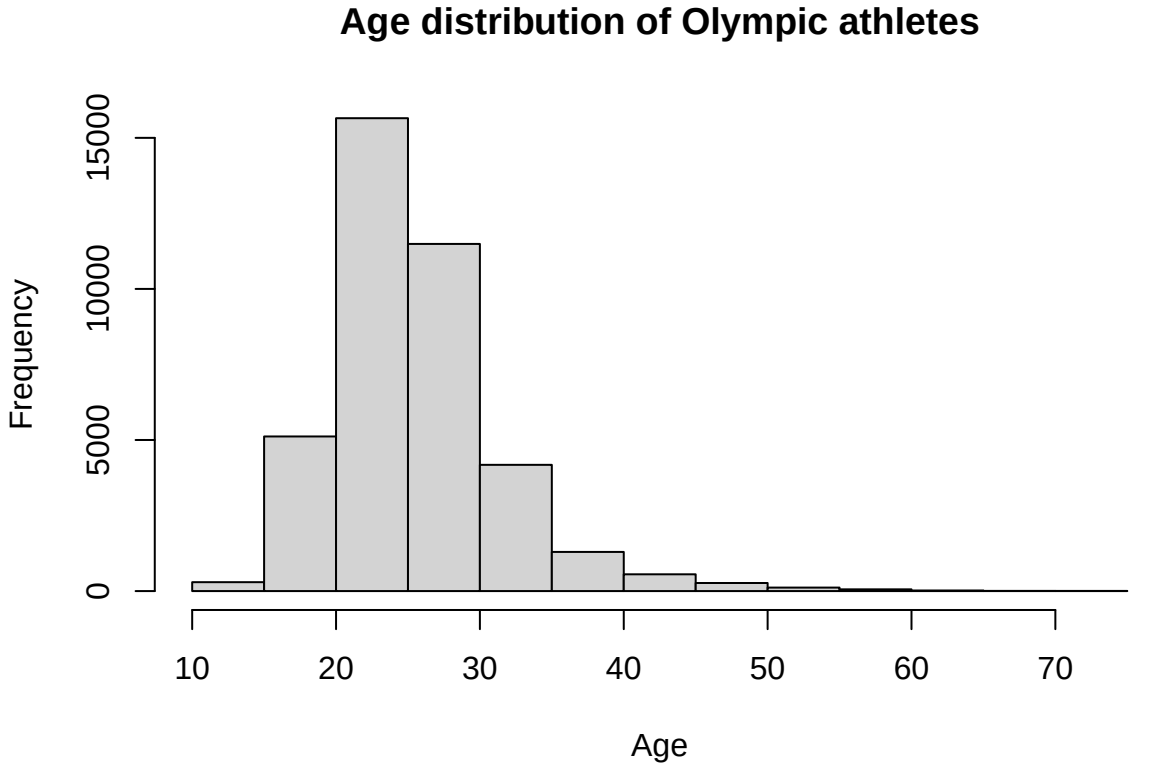



```
#For each year, there is a general positive trend in the graph. This means that for each year, the high
```

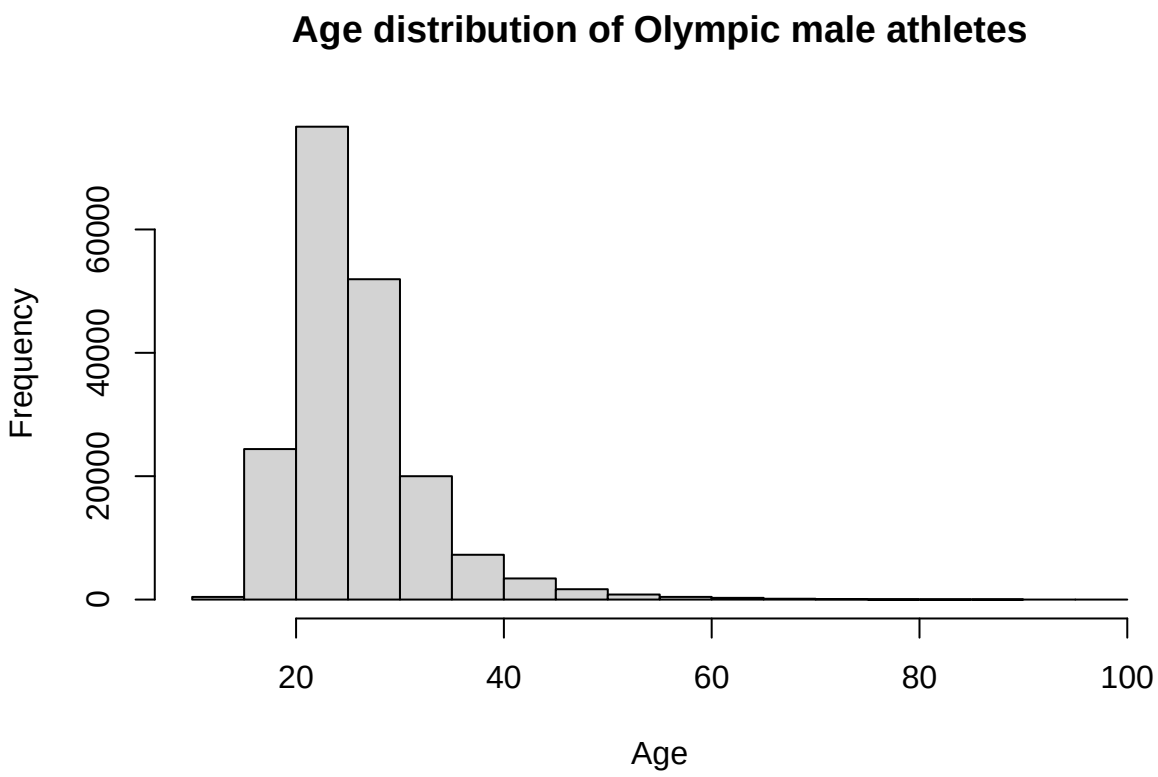
```
#Question 3
#Is the distribution of medals one similar to the overall age distribution?
#For this data, we are taking data from all years of the olympics
#Below is the age distribution of all athletes in this dataset
hist(athlete_events$Age)
```



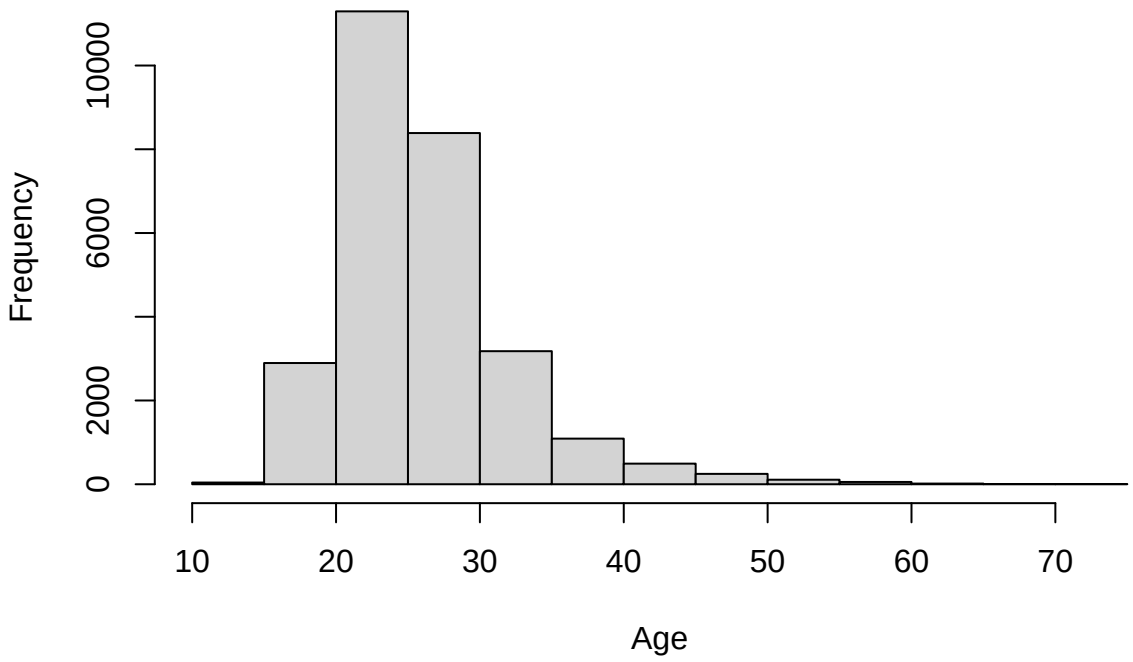

```r
#we filter out only medal winners
only.medals <- filter(athlete_events, Medal == "Bronze" | Medal == "Silver" | Medal == "Gold")

#Here is the age distribution among medal winners
hist(only.medals$Age, main = "Age distribution of Olympic athletes", xlab = "Age",ylab = "Frequency")
```



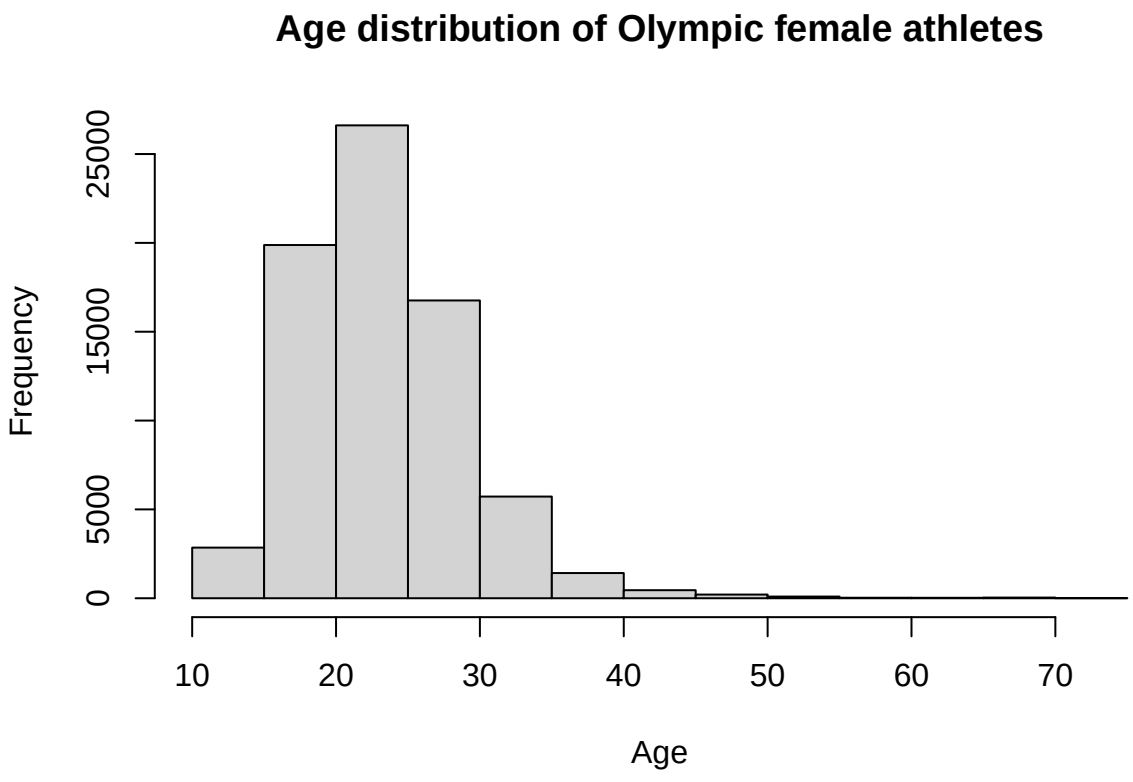
```r
?hist()
#Comparing the two histograms, it appears that the age distribution for medal winners is proportional t

#I will further look to see if there are differences for age distribution of just males and females.

#First we filter accordingly
only.male <- filter(athlete_events, Sex == "M")

only.male.medal <- filter(only.medals, Sex == "M")

hist(only.male$Age, main = "Age distribution of Olympic male athletes", xlab = "Age",ylab = "Frequency"
```



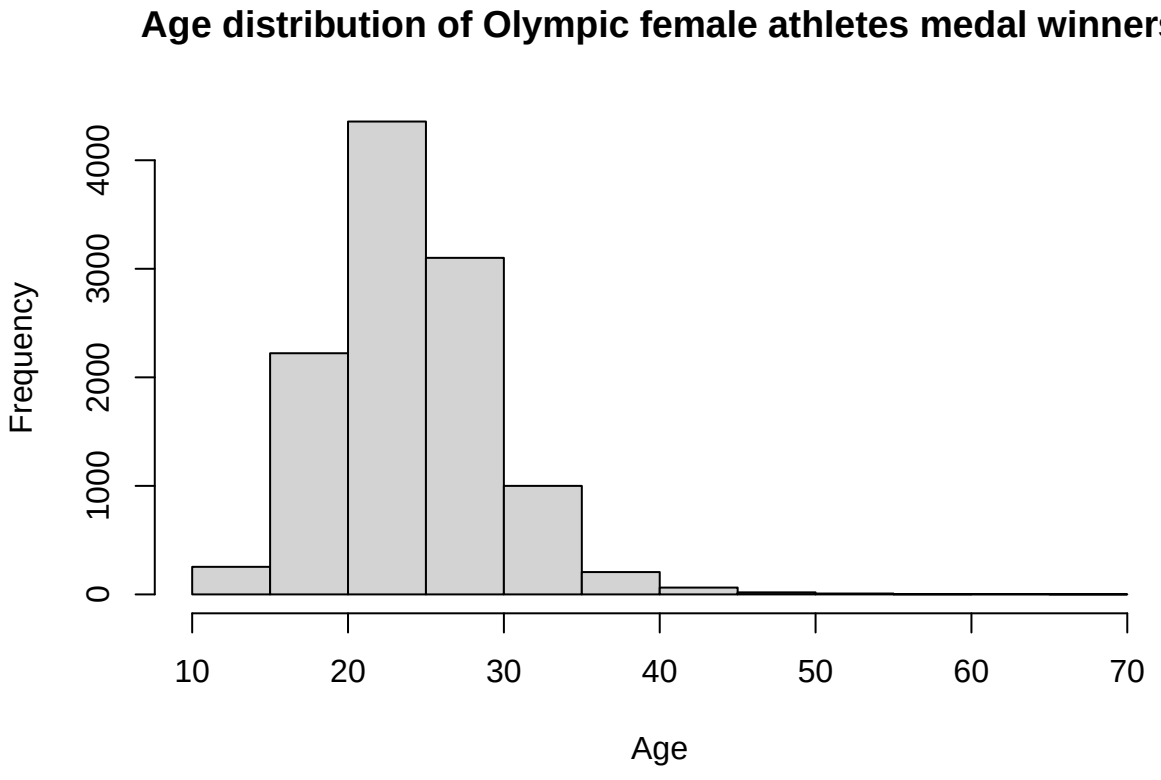

```r
hist(only.male.medal$Age, main = "Age distribution of Olympic male athletes medal winners", xlab = "Age
```




```
#For male athletes, the age distribution of medal winners appears to be very close to the overall age d

only.female <- filter(athlete_events, Sex == "F")

only.female.medal <- filter(only.medals, Sex == "F")

hist(only.female$Age, main = "Age distribution of Olympic female athletes", xlab = "Age",ylab = "Freque
```




```
hist(only.female.medal$Age, main = "Age distribution of Olympic female athletes medal winners", xlab =
```




```
#For females, it appears that the medal winners occur at a higher rate compared to the total proportion
```