Assignment4

Load worldrecord.csv and answer the following?

library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

dataset<-read.csv("WorldRecords.csv")

1. How many different types of events (e.g. “Mens 100m”, “Womens shotput” etc)  
   are represented in the dataset.

unique(dataset$Event)

## [1] Mens 100m Womens 100m Mens 800m Womens 800m   
## [5] Mens TripleJump Mens Mile Womens Mile Mens Polevault   
## [9] Mens Shotput Womens Shotput   
## 10 Levels: Mens 100m Mens 800m Mens Mile Mens Polevault ... Womens Shotput

1. In what year did Usain Bolt first break the world record for the Men’s 100m?

head(dataset[dataset$Athlete=='Usain Bolt',"Year"],n=1)

## [1] 2008

1. Which variable tells us the record setting time or distance?  
   The variable name in the data set is? What type of the variable is this?

Type variable tells us the record setting time or distance.

#Type of the variable  
 str(dataset$Type)

## Factor w/ 2 levels "distance","time": 2 2 2 2 2 2 2 2 2 2 ...

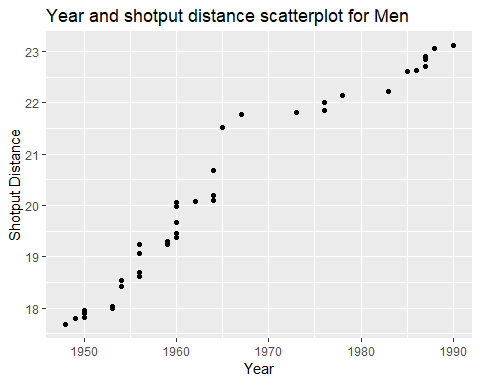
1. Create a subset of the dataset that contains only the world record cases for men’s shotput and women’s shotput.

sub1=subset(dataset,Event =='Mens Shotput' | Event =='Womens Shotput')  
head(sub1)

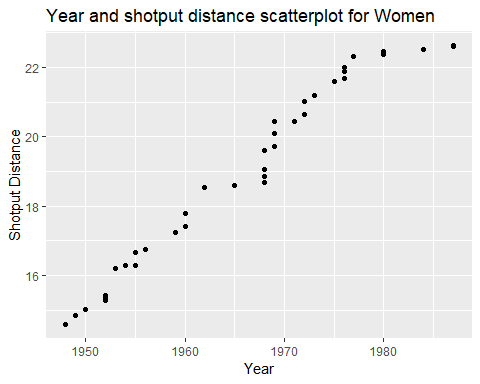
## Event Type Record Athlete Nationality  
## 206 Mens Shotput distance 17.68 ÊCharlie FonvilleÊ USA  
## 207 Mens Shotput distance 17.79 ÊJim FuchsÊ USA  
## 208 Mens Shotput distance 17.82 ÊJim FuchsÊ USA  
## 209 Mens Shotput distance 17.90 ÊJim FuchsÊ USA  
## 210 Mens Shotput distance 17.95 ÊJim FuchsÊ USA  
## 211 Mens Shotput distance 18.00 ÊParry O'BrienÊ USA  
## Location Year  
## 206 Lawrence, U.S. 1948  
## 207 Oslo, Norway 1949  
## 208 Los Angeles, U.S. 1950  
## 209 Visby, Sweden 1950  
## 210 Eskilstuna, Sweden 1950  
## 211 Fresno, U.S. 1953

1. Create a scatter plot of the year and record shotput distance one for men and one for women.

library(ggplot2)  
shotput\_men<-subset(dataset,Event =='Mens Shotput')  
ggplot(shotput\_men,aes(x=Year,y=Record))+geom\_point()+ggtitle("Year and shotput distance scatterplot for Men")+ xlab("Year")+ylab("Shotput Distance")



shotput\_women<-subset(dataset,Event =='Womens Shotput')  
  
ggplot(shotput\_women,aes(x=Year,y=Record))+geom\_point()+ggtitle("Year and shotput distance scatterplot for Women")+ xlab("Year")+ylab("Shotput Distance")



1. Find the average/mean time for each event.How many athletes have time more  
   than average in each event.

grouped\_by\_event<-group\_by(dataset[dataset$Type=='time',],Event)  
mean\_event<-summarize(grouped\_by\_event,avg\_time=mean(Record))  
mean\_event

## # A tibble: 6 x 2  
## Event avg\_time  
## <fct> <dbl>  
## 1 Mens 100m 9.85  
## 2 Mens 800m 105.   
## 3 Mens Mile 237.   
## 4 Womens 100m 10.9   
## 5 Womens 800m 128.   
## 6 Womens Mile 264.

finaldatset<-data.frame()  
for(i in 1:nrow(mean\_event)){  
 finaldatset<-dataset[dataset$Event==mean\_event$Event[i] & dataset$Record>mean\_event$avg\_time[i],]  
 print(unique(finaldatset$Event))  
 print(nrow(finaldatset))  
}

## [1] Mens 100m  
## 10 Levels: Mens 100m Mens 800m Mens Mile Mens Polevault ... Womens Shotput  
## [1] 9  
## [1] Mens 800m  
## 10 Levels: Mens 100m Mens 800m Mens Mile Mens Polevault ... Womens Shotput  
## [1] 10  
## [1] Mens Mile  
## 10 Levels: Mens 100m Mens 800m Mens Mile Mens Polevault ... Womens Shotput  
## [1] 15  
## [1] Womens 100m  
## 10 Levels: Mens 100m Mens 800m Mens Mile Mens Polevault ... Womens Shotput  
## [1] 4  
## [1] Womens 800m  
## 10 Levels: Mens 100m Mens 800m Mens Mile Mens Polevault ... Womens Shotput  
## [1] 13  
## [1] Womens Mile  
## 10 Levels: Mens 100m Mens 800m Mens Mile Mens Polevault ... Womens Shotput  
## [1] 5

1. Select the athlete who took most time in men’s 100m and women’s event.

top\_n(dataset[dataset$Event=='Mens 100m',],1,Record)

## Event Type Record Athlete Nationality Location Year  
## 1 Mens 100m time 10.06 Bob Hayes United States Tokyo, Japan 1964

top\_n(dataset[dataset$Event=='Womens 100m',],1,Record)

## Event Type Record Athlete Nationality  
## 1 Womens 100m time 11.07 Wyomia Tyus ÊUnited States  
## 2 Womens 100m time 11.07 Renate Stecher ÊEast Germany  
## Location Year  
## 1 Mexico City, Mexico 1968  
## 2 Munich, West Germany 1972

1. Which country won maximum times of men’s 100m event?

country\_group<-group\_by(dataset[dataset$Event=='Mens 100m',],Nationality)  
dataset2<-summarize(country\_group,country\_count=n())  
head(dataset2[order(-dataset2$country\_count),],n=1)

## # A tibble: 1 x 2  
## Nationality country\_count  
## <fct> <int>  
## 1 United States 11

1. How many athletes are there in each event?

group\_event<-group\_by(dataset,Event)  
summarize(group\_event,athlete\_count=n\_distinct(Athlete))

## # A tibble: 10 x 2  
## Event athlete\_count  
## <fct> <int>  
## 1 Mens 100m 11  
## 2 Mens 800m 18  
## 3 Mens Mile 23  
## 4 Mens Polevault 22  
## 5 Mens Shotput 13  
## 6 Mens TripleJump 18  
## 7 Womens 100m 8  
## 8 Womens 800m 21  
## 9 Womens Mile 10  
## 10 Womens Shotput 12

1. which country has maximum wins?

country\_group<-group\_by(dataset,Nationality)  
dataset<-summarize(country\_group,country\_count=n())  
head(dataset[order(-dataset$country\_count),],n=1)

## # A tibble: 1 x 2  
## Nationality country\_count  
## <fct> <int>  
## 1 URS 49