Lab 1

source("http://www.openintro.org/stat/data/arbuthnot.R")

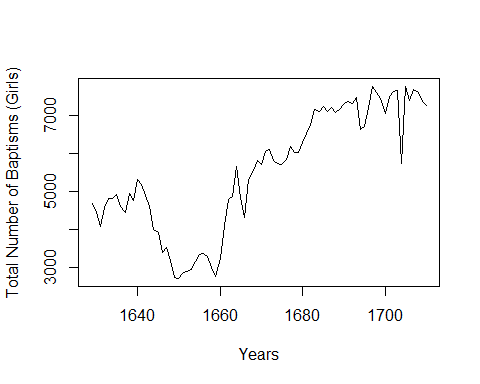
## Exercise 1: What caommand would you use to extract just the counts of girls baptized?

arbuthnot$girls

## [1] 4683 4457 4102 4590 4839 4820 4928 4605 4457 4952 4784 5332 5200 4910  
## [15] 4617 3997 3919 3395 3536 3181 2746 2722 2840 2908 2959 3179 3349 3382  
## [29] 3289 3013 2781 3247 4107 4803 4881 5681 4858 4319 5322 5560 5829 5719  
## [43] 6061 6120 5822 5738 5717 5847 6203 6033 6041 6299 6533 6744 7158 7127  
## [57] 7246 7119 7214 7101 7167 7302 7392 7316 7483 6647 6713 7229 7767 7626  
## [71] 7452 7061 7514 7656 7683 5738 7779 7417 7687 7623 7380 7288

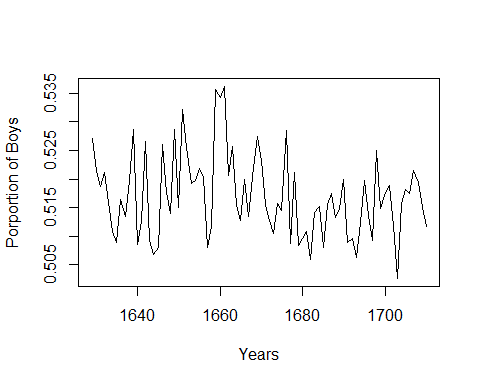
## Exercise 2: Use the information in the help file to add a title to your plot and to give the x- and y- axes more appropriate names.

plot(x = arbuthnot$year, y = arbuthnot$girls, type = "l", ylab = "Total Number of Baptisms (Girls)", xlab = "Years")



## Make a plot of the proportion of boys over time

plot(arbuthnot$year, arbuthnot$boys / (arbuthnot$boys + arbuthnot$girls), type = "l",ylab = "Porportion of Boys", xlab = "Years")



source("http://www.openintro.org/stat/data/present.R")

## 1.What years are included in this data set? What are the dimensions of the data frame, and what are the variable (i.e., column) names?

dim(present)

## [1] 63 3

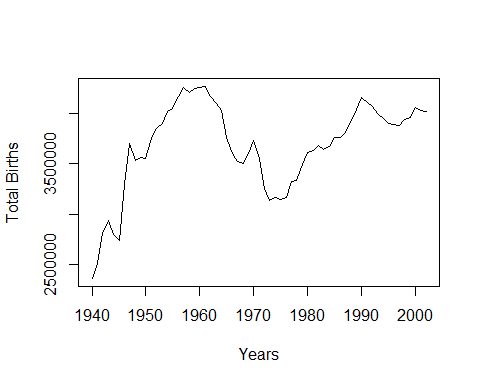
names(present)

## [1] "year" "boys" "girls"

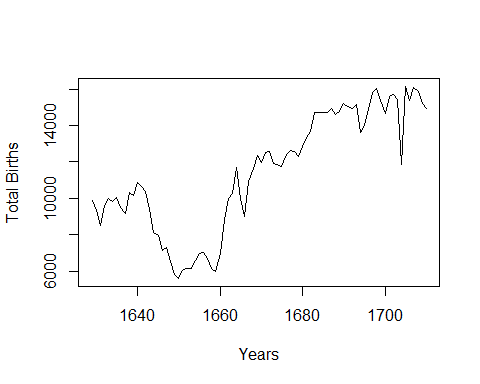
## There are 63 rows of data with 3 column names. The column names are “year”, “boys”, and “girls”.

## 2.How do the present-day total birth counts per year compare to Arbuthnot’s? Are they on a similar scale? Explain your reasoning

plot(present$year, present$boys + present$girls, type ="l",ylab = "Total Births", xlab = "Years")



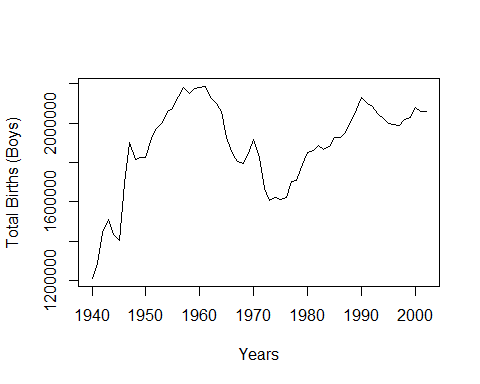
plot(arbuthnot$year, arbuthnot$boys + arbuthnot$girls, type ="l", ylab = "Total Births", xlab = "Years")



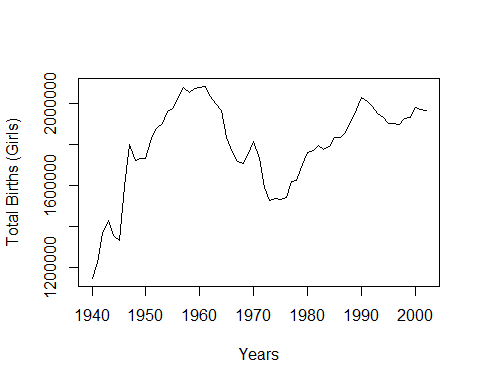
# There are more births per year in the present-day data than the Arbuthnot’s data. There is an early drop off in Arbuthnots but not in the present day. While both sets of data seem to recover as time goes on, the data sets are not similar.

## 3. Does Arbuthnot’s observation about boys being born in greater proportion than girls hold up in the U.S. during this time period? Explain.

plot(present$year, present$boys, type ="l",ylab = "Total Births (Boys)", xlab = "Years")



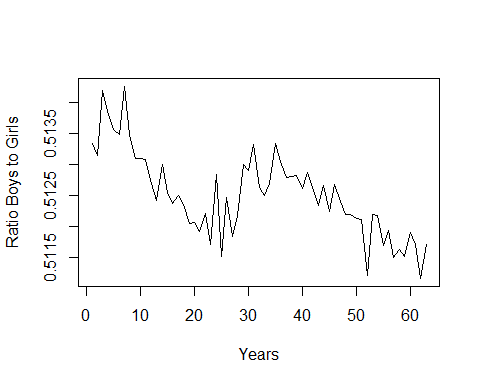
plot(present$year, present$girls, type ="l",ylab = "Total Births (Girls)", xlab = "Years")



## No, Arbuthnot’s obercation about boys being born in greater numbers does not hold up in the U.S during this time periosd. The amounts are about equal.

## 4. Make a plot that displays the boy-to-girl ratio for every year in the data set. Give the plot an informative title, and label the x- and y-axes with useful titles. What trend do you see? Export your plot and include it with your answer to this question.

plot(present$boys/(present$boys + present$girls), type ="l",ylab = "Ratio Boys to Girls", xlab = "Years")

 ##As the year increases, the ratio decreases

## Find and report the 5-number summary

quantile(present$boys)

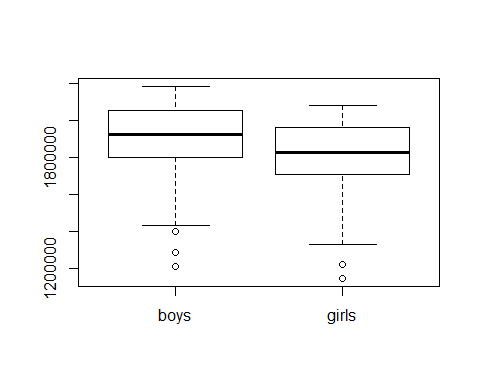
## 0% 25% 50% 75% 100%   
## 1211684 1799857 1924868 2058524 2186274

quantile(present$girls)

## 0% 25% 50% 75% 100%   
## 1148715 1711405 1831679 1965538 2082052

## Looking at your box plot, you will see that there are three outliers present. Find the values for those outliers, and then do the fence calculations needed to classify each outlier as a mild low, extreme low, mild high, or extreme high outlier. Show and label your calculations, and explain your reasoning.

boxplot(present[,-1])



IQR(present$girls)

## [1] 254133

quantile(present$girls,.25) - (1.5 \* IQR(present$girls))

## 25%   
## 1330205

quantile(present$girls,.75) + (1.5 \* IQR(present$girls))

## 75%   
## 2346737

quantile(present$girls,.25) - (3 \* IQR(present$girls))

## 25%   
## 949005.5

quantile(present$girls,.75) + (3 \* IQR(present$girls))

## 75%   
## 2727937

IQR(present$boys)

## [1] 258666.5

quantile(present$boys,.25) - (1.5 \* IQR(present$boys))

## 25%   
## 1411857

quantile(present$boys,.75) + (1.5 \* IQR(present$boys))

## 75%   
## 2446523

quantile(present$boys,.25) - (3 \* IQR(present$boys))

## 25%   
## 1023858

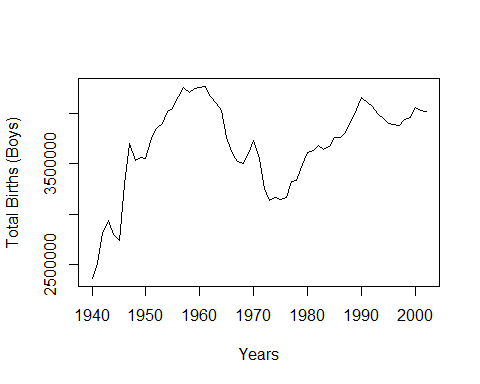
quantile(present$boys,.75) + (3 \* IQR(present$boys))

## 75%   
## 2834523

# The outlier are extreme low.

## In what year did we see the lowest total number of births in the U.S.? Include a screenshot of your workspace with comments that explain how you arrived at the answer to this question.

plot(present$year, present$boys+present$girls, type ="l",ylab = "Total Births (Boys)", xlab = "Years")

 #1973

## Generate a stem-and-leaf plot of the total U.S. births using scale = 2. Explain what value 27 | 9 represents in the plot. How does this value compare to the actual birth count that it represents? Also, describe the shape of the distribution. Include a screenshot of your stem-and-leaf plot.

stem(present$boys + present$girls, scale=2)

##   
## The decimal point is 5 digit(s) to the right of the |  
##   
## 23 | 6  
## 24 |   
## 25 | 1  
## 26 |   
## 27 | 49  
## 28 | 1  
## 29 | 4  
## 30 |   
## 31 | 4467  
## 32 | 69  
## 33 | 33  
## 34 | 9  
## 35 | 024566  
## 36 | 0113478  
## 37 | 035666  
## 38 | 1589  
## 39 | 001456  
## 40 | 022334567  
## 41 | 01667  
## 42 | 04567

# The plot is left skewed. 27|9 means that there was a number that began with 279xxxxx. Its not exact number because a lot of the numbers have been dropped.