Introduction to R and RStudio

Tom Buonora

Dr. Arbuthnot's Baptism Records

To get started, let's take a peek at the data.

```
source('C:/Users/arono/Documents/R/win-library/4.0/DATA606/labs/Lab1/more/arbuthnot.r')
```

```
arbuthnot
```

```
##    year boys girls
## 1  1629 5218  4683
## 2  1630 4858  4457
## 3  1631 4422  4102
## 4  1632 4994  4590
## 5  1633 5158  4839
## 6  1634 5035  4820
## 7  1635 5106  4928
## 8  1636 4917  4605
## 9  1637 4703  4457
## 10 1638 5359  4952
## 11 1639 5366  4784
## 12 1640 5518  5332
## 13 1641 5470  5200
## 14 1642 5460  4910
## 15 1643 4793  4617
## 16 1644 4107  3997
## 17 1645 4047  3919
## 18 1646 3768  3395
## 19 1647 3796  3536
## 20 1648 3363  3181
## 21 1649 3079  2746
## 22 1650 2890  2722
## 23 1651 3231  2840
## 24 1652 3220  2908
## 25 1653 3196  2959
## 26 1654 3441  3179
## 27 1655 3655  3349
## 28 1656 3668  3382
## 29 1657 3396  3289
## 30 1658 3157  3013
## 31 1659 3209  2781
## 32 1660 3724  3247
## 33 1661 4748  4107
## 34 1662 5216  4803
```

```
## 35 1663 5411  4881
## 36 1664 6041  5681
## 37 1665 5114  4858
## 38 1666 4678  4319
## 39 1667 5616  5322
## 40 1668 6073  5560
## 41 1669 6506  5829
## 42 1670 6278  5719
## 43 1671 6449  6061
## 44 1672 6443  6120
## 45 1673 6073  5822
## 46 1674 6113  5738
## 47 1675 6058  5717
## 48 1676 6552  5847
## 49 1677 6423  6203
## 50 1678 6568  6033
## 51 1679 6247  6041
## 52 1680 6548  6299
## 53 1681 6822  6533
## 54 1682 6909  6744
## 55 1683 7577  7158
## 56 1684 7575  7127
## 57 1685 7484  7246
## 58 1686 7575  7119
## 59 1687 7737  7214
## 60 1688 7487  7101
## 61 1689 7604  7167
## 62 1690 7909  7302
## 63 1691 7662  7392
## 64 1692 7602  7316
## 65 1693 7676  7483
## 66 1694 6985  6647
## 67 1695 7263  6713
## 68 1696 7632  7229
## 69 1697 8062  7767
## 70 1698 8426  7626
## 71 1699 7911  7452
## 72 1700 7578  7061
## 73 1701 8102  7514
## 74 1702 8031  7656
## 75 1703 7765  7683
## 76 1704 6113  5738
## 77 1705 8366  7779
## 78 1706 7952  7417
## 79 1707 8379  7687
## 80 1708 8239  7623
## 81 1709 7840  7380
## 82 1710 7640  7288
```

```
# apparently glimpse has been imported through 2 packages, tibble and dplyr
glimpse(arbuthnot)
```

```
## Rows: 82
## Columns: 3
```

```
## $ year  <int> 1629, 1630, 1631, 1632, 1633, 1634, 1635, 1636, 1637, 1638, 1...
## $ boys  <int> 5218, 4858, 4422, 4994, 5158, 5035, 5106, 4917, 4703, 5359, 5...
## $ girls <int> 4683, 4457, 4102, 4590, 4839, 4820, 4928, 4605, 4457, 4952, 4...
```

```
glimpse(arbuthnot)
```

```
## Rows: 82
## Columns: 3
## $ year  <int> 1629, 1630, 1631, 1632, 1633, 1634, 1635, 1636, 1637, 1638, 1...
## $ boys  <int> 5218, 4858, 4422, 4994, 5158, 5035, 5106, 4917, 4703, 5359, 5...
## $ girls <int> 4683, 4457, 4102, 4590, 4839, 4820, 4928, 4605, 4457, 4952, 4...
```

Some Exploration

```
head(arbuthnot$boys)
```

```
## [1] 5218 4858 4422 4994 5158 5035
```

1. What command would you use to extract just the counts of girls baptized? Try it!

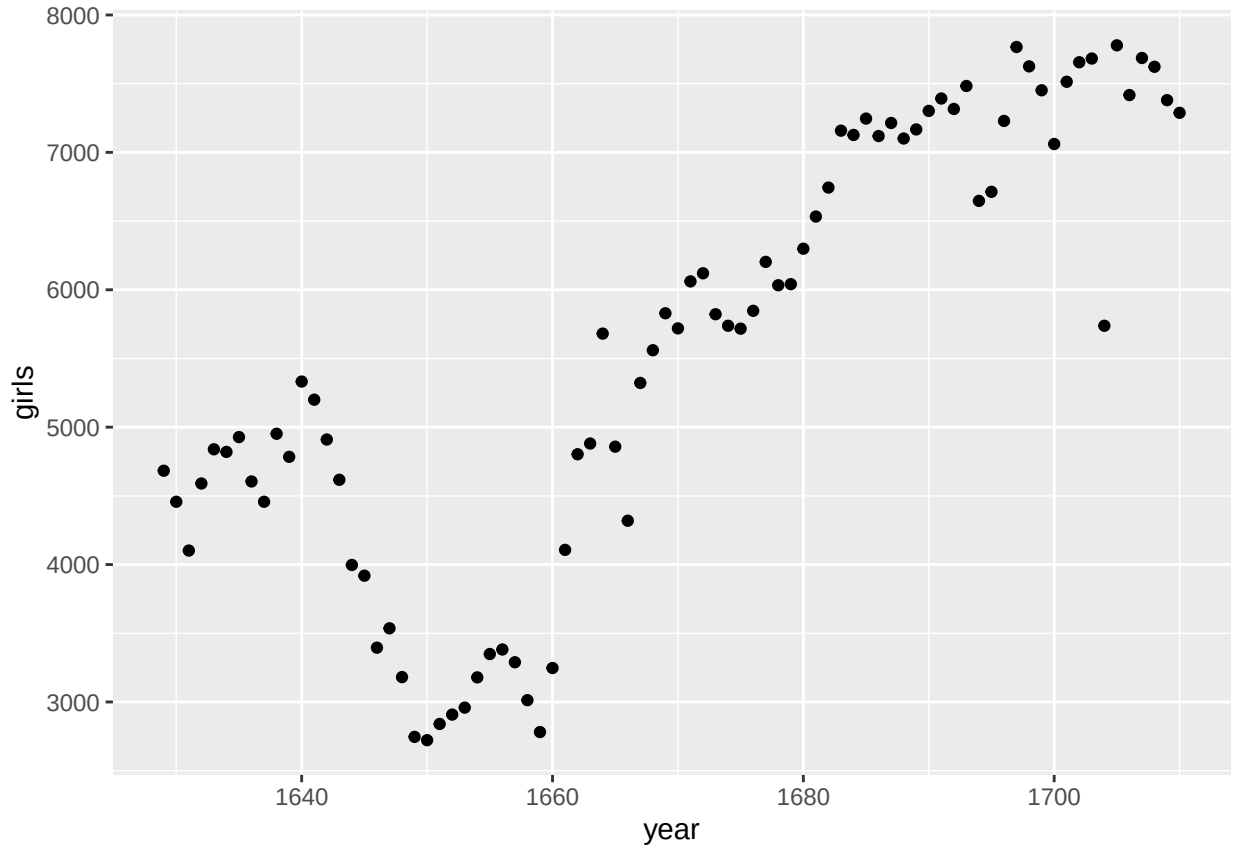
```
head(arbuthnot$girls)
```

```
## [1] 4683 4457 4102 4590 4839 4820
```

Data visualization

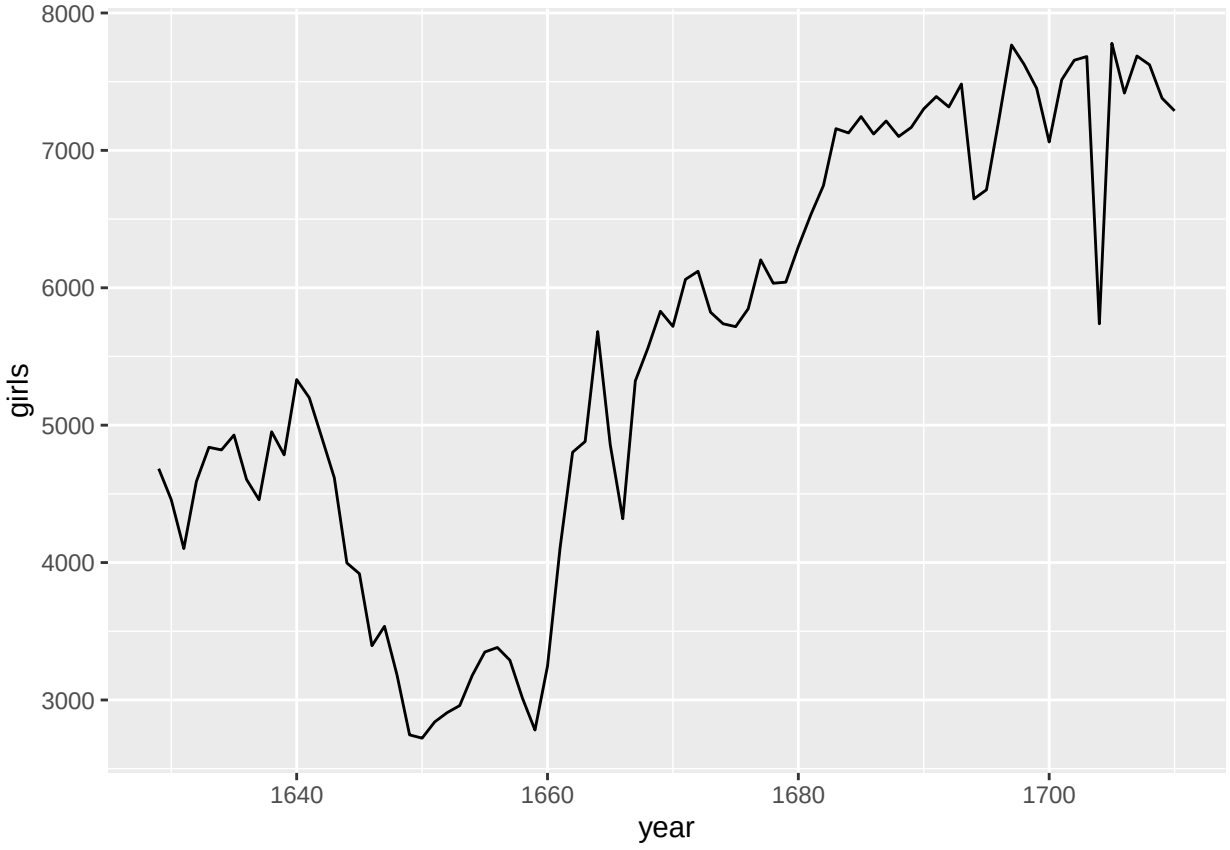
R has some powerful functions for making graphics. We can create a simple plot of the number of girls baptized per year with the command

```
ggplot(data = arbuthnot, aes(x = year, y = girls)) +  geom_point()
```



Replace `geom_point()` with `geom_line()`.

```
ggplot(data = arbuthnot, aes(x = year, y = girls)) +
  geom_line()
```



1. Is there an apparent trend in the number of girls baptized over the years? How would you describe it? (To ensure that your lab report is comprehensive, be sure to include the code needed to make the plot as well as your written interpretation.)

The overall trend shows an increase in the number of girls getting baptised each year. But there were some dramatic drops as well around 1640 and 1720, and some dramatic increaeses around 1660 and 1722.

The below chunk calculates the delta of girls baptised each year and then prints out the top 5 years of increase and decrease.

```
arbuthnot <- arbuthnot %>%
  mutate(delta_girls = 0)

for (i in 2:nrow(arbuthnot))
{
  if (i==2)
  {
    prev_girls=arbuthnot[1,"girls"]
  }



  current_girls = arbuthnot[i,"girls"]
  delta_girls = current_girls-prev_girls
  arbuthnot[i,"delta_girls"]<-delta_girls
```

```
  prev_girls=current_girls
}

arbuthnot[which.max(arbuthnot$delta_girls),]
```

```
##    year boys girls delta_girls
## 77 1705 8366  7779        2041
```

```
arbuthnot[which.min(arbuthnot$delta_girls),]
```

```
##    year boys girls delta_girls
## 76 1704 6113  5738       -1945
```

```
arbuthnot_sorted<-arbuthnot %>% arrange(desc(delta_girls))
print ("The greatest increaseses : ")
```

```
## [1] "The greatest increaseses : "
```

```
head(arbuthnot_sorted)
```

```
##   year boys girls delta_girls
## 1 1705 8366  7779        2041
## 2 1667 5616  5322        1003
## 3 1661 4748  4107         860
## 4 1664 6041  5681         800
## 5 1662 5216  4803         696
## 6 1640 5518  5332         548
```

```
print("The greatest decreases")
```

```
## [1] "The greatest decreases"
```

```
tail(arbuthnot_sorted)
```

```
##    year boys girls delta_girls
## 77 1646 3768  3395        -524
## 78 1666 4678  4319        -539
## 79 1644 4107  3997        -620
## 80 1665 5114  4858        -823
## 81 1694 6985  6647        -836
## 82 1704 6113  5738       -1945
```

I was close. It was actually 1704 and 1705 which had the most dramatic deltas.

Display rows where girls are greater than boys

```
nrow(subset(arbuthnot, girls > boys))
```

```
## [1] 0
```

Now display rows where boys are greater than girls

```
nrow(subset(arbuthnot, boys > girls))
```

```
## [1] 82
```

I would like to verify if we should assume that this ratio in baptism records actually infers the same ratio in actual births.

R as a big calculator

If we add the vector for baptisms for boys to that of girls, R will compute all sums simultaneously.

```
arbuthnot$boys + arbuthnot$girls
```

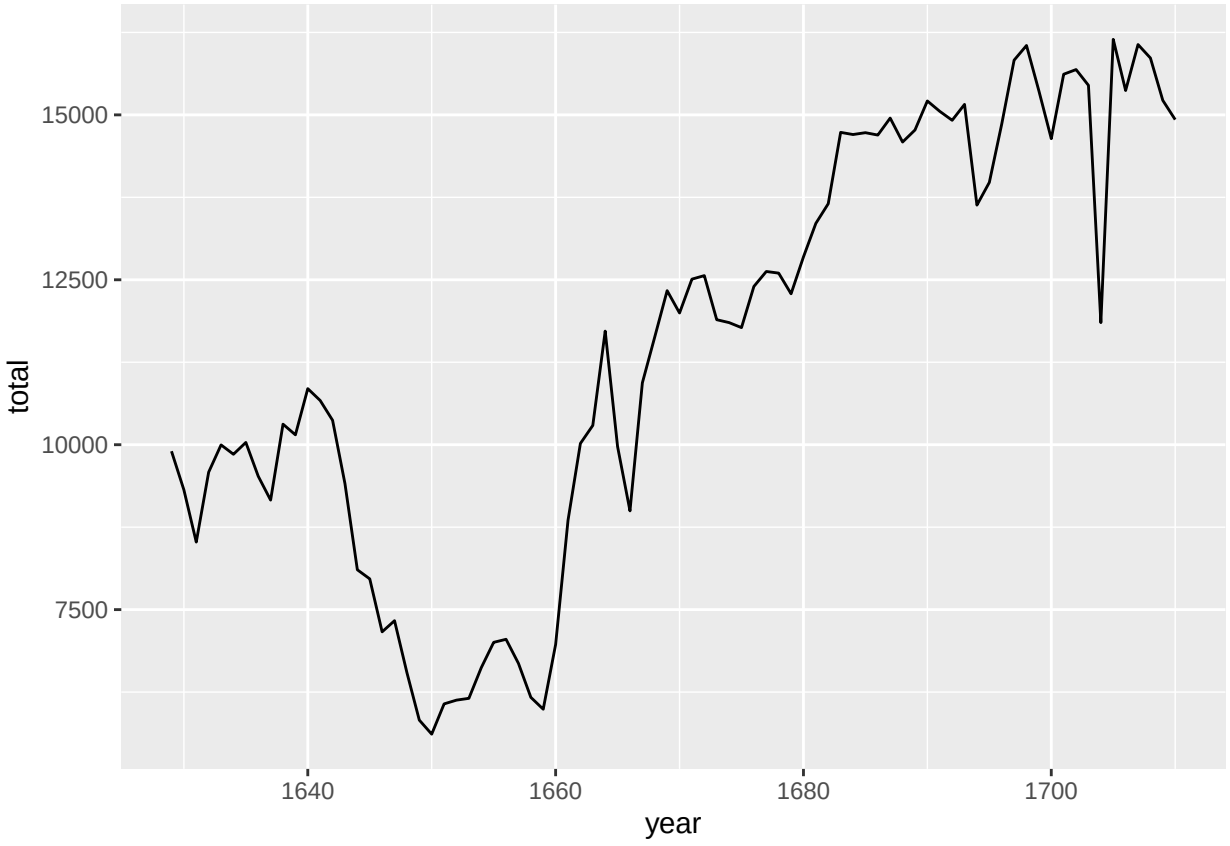
```
##  [1]  9901  9315  8524  9584  9997  9855 10034  9522  9160 10311 10150 10850
## [13] 10670 10370  9410  8104  7966  7163  7332  6544  5825  5612  6071  6128
## [25]  6155  6620  7004  7050  6685  6170  5990  6971  8855 10019 10292 11722
## [37]  9972  8997 10938 11633 12335 11997 12510 12563 11895 11851 11775 12399
## [49] 12626 12601 12288 12847 13355 13653 14735 14702 14730 14694 14951 14588
## [61] 14771 15211 15054 14918 15159 13632 13976 14861 15829 16052 15363 14639
## [73] 15616 15687 15448 11851 16145 15369 16066 15862 15220 14928
```

Adding a new variable to the data frame

We'll be using this new vector to generate some plots, so we'll want to save it as a permanent column in our data frame.

```
arbuthnot <- arbuthnot %>%
  mutate(total = boys + girls)
```

```
ggplot(data = arbuthnot, aes(x = year, y = total)) +
  geom_line()
```



Create a boy to girl ratio

```
arbuthnot <- arbuthnot %>%
  mutate(boy_to_girl_ratio = boys / girls)
```

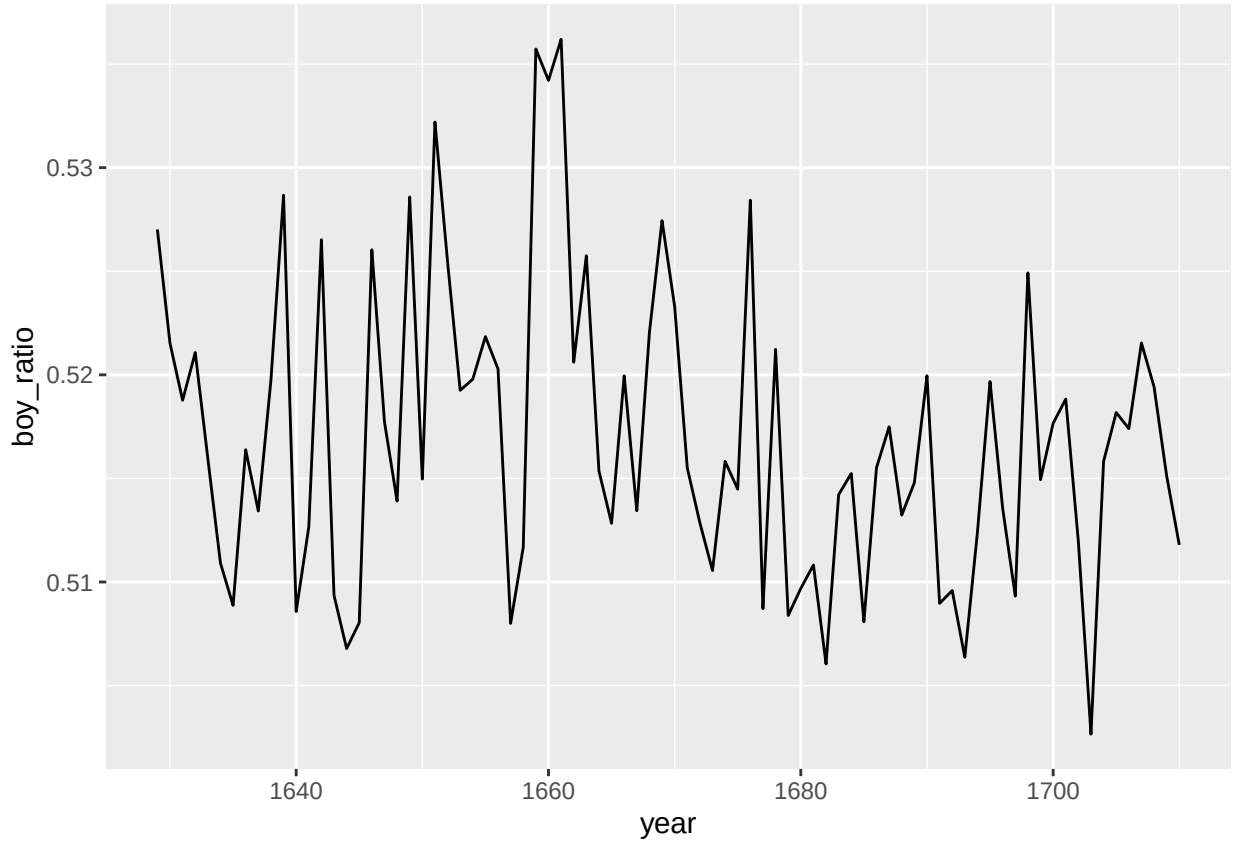
Create a boy to total ratio

```
arbuthnot <- arbuthnot %>%
  mutate(boy_ratio = boys / total)
```

Note **mutate()** *adds new variables and preserves existing ones.* **transmute()** *adds new variables and drops existing ones.*

3. Now, generate a plot of the proportion of boys born over time. What do you see?

```
ggplot(data = arbuthnot, aes(x = year, y = boy_ratio)) +
  geom_line()
```



My comment is that the graph seems choppy, but only because the scale is so narrow. I would characterize the male to female birth ratio as consistent within the 51 to 53 percent range.

```
arbuthnot <- arbuthnot %>%
  mutate(more_boys = boys > girls)
```

More Practice

```
arbuthnot %>%
  summarize(min = min(boys), max = max(boys))
```

```
##    min  max
## 1 2890 8426
```

1. What years are included in this data set? What are the dimensions of the data frame? What are the variable (column) names?

Since there are 82 rows, I will just print the minimum/maximum of year

```
# arbuthnot["year"]


present <- present %>%  mutate(total = boys + girls)
```

```
present <- present %>%  mutate(boy_to_girl_ratio = boys / girls)
present <- present %>% mutate(more_boys = boys > girls)


sprintf("This study spanned the years from %s to  %s ", min(present["year"]),max(present["year"]))
```

```
## [1] "This study spanned the years from 1940 to  2002 "
```

```
d_df <- dim(present)
sprintf("The dimensions of the table are %d rows and %d columns ", d_df[1],d_df[2])
```

```
## [1] "The dimensions of the table are 63 rows and 6 columns "
```

```
cnames<-as.data.frame(names(present))
knitr::kable(cnames, caption='Column Names',col.names = "")
```

Table 1: Column Names

year
boys
girls
total
boy_to_girl_ratio
more_boys

1. How do these counts compare to Arbuthnot's? Are they of a similar magnitude?

The counts are signficantly higher. The actual numbers will be shown in a little bit when I calculate the mean totals

```
present_range<- max(present["year"]) - min(present["year"])

arbuthnot_range<- max(arbuthnot["year"]) - min(arbuthnot["year"])

sprintf("The arbuthnot years were %s to  %s or %d total years",
        min(arbuthnot["year"]),max(arbuthnot["year"]), arbuthnot_range)
```

```
## [1] "The arbuthnot years were 1629 to  1710 or 81 total years"
```

```
sprintf("Meanwhile the present study covered %d years",present_range)
```

```
## [1] "Meanwhile the present study covered 62 years"
```

```
present_avg_total<-format(mean(present$total),big.mark = ",")

arbuthnot_avg_total<-format(mean(arbuthnot$total),big.mark = ",")

sprintf("The average total count was %s for arbuthnot and  %s for present ",
          arbuthnot_avg_total, present_avg_total)
```

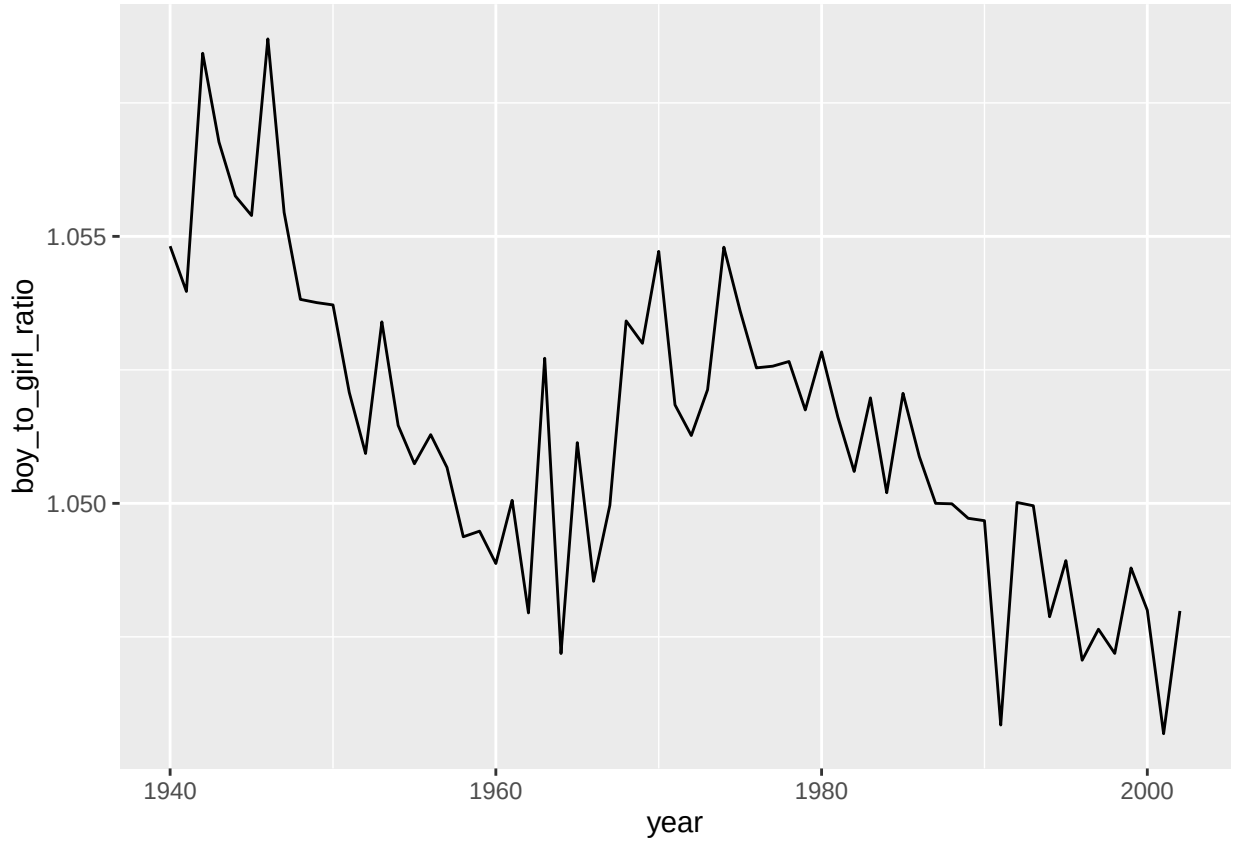
```
## [1] "The average total count was 11,441.74 for arbuthnot and  3,679,515 for present "
```

```
sprintf("The average ration was %.4f for arbuthnot and  %.4f for present ",
          mean(arbuthnot$boy_to_girl_ratio),  mean(present$boy_to_girl_ratio))
```

```
## [1] "The average ration was 1.0707 for arbuthnot and  1.0514 for present "
```

1. Make a plot that displays the proportion of boys born over time. What do you see? Does Arbuthnot's observation about boys being born in greater proportion than girls hold up in the U.S.?

```
ggplot(data = present, aes(x = year, y = boy_to_girl_ratio)) +  geom_line()
```



I see the same ratio which, to me, is strong evidence that the ratio is natural to our species.

1. In what year did we see the most total number of births in the U.S.?

```
# present_totals<-present %>% arrange(desc(total))

present[which.max(present$total),]
```

```
## # A tibble: 1 x 6
##    year    boys   girls   total boy_to_girl_ratio more_boys
##   <dbl>   <dbl>   <dbl>   <dbl>             <dbl> <lgl>
## 1  1961 2186274 2082052 4268326              1.05 TRUE
```