DATA 606, Lab 0 - Introduction to R and RStudio

Kavya Beheraj

February 4, 2018

This lab is an introduction to R and RStudio.

The datasets I used are:

- Dr. Arbuthnot's Baptism Records
- US Birth Records (Present), compiled by OpenIntro Statistics and recorded by the Centers for Disease Control

---

The Data: Dr. Arbuthnot's Baptism Records

```
arbuthnot = read.csv("C:/Users/Kavya/Desktop/Education/MS Data Science/DATA 606 (Statistics and Probabil

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
```

```
dim(arbuthnot)
```

```
## [1] 82  3
```

```
names(arbuthnot)
```

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

Some Exploration

```
arbuthnot$boys
```

```
##  [1] 5218 4858 4422 4994 5158 5035 5106 4917 4703 5359 5366 5518 5470 5460
## [15] 4793 4107 4047 3768 3796 3363 3079 2890 3231 3220 3196 3441 3655 3668
## [29] 3396 3157 3209 3724 4748 5216 5411 6041 5114 4678 5616 6073 6506 6278
## [43] 6449 6443 6073 6113 6058 6552 6423 6568 6247 6548 6822 6909 7577 7575
## [57] 7484 7575 7737 7487 7604 7909 7662 7602 7676 6985 7263 7632 8062 8426
## [71] 7911 7578 8102 8031 7765 6113 8366 7952 8379 8239 7840 7640
```

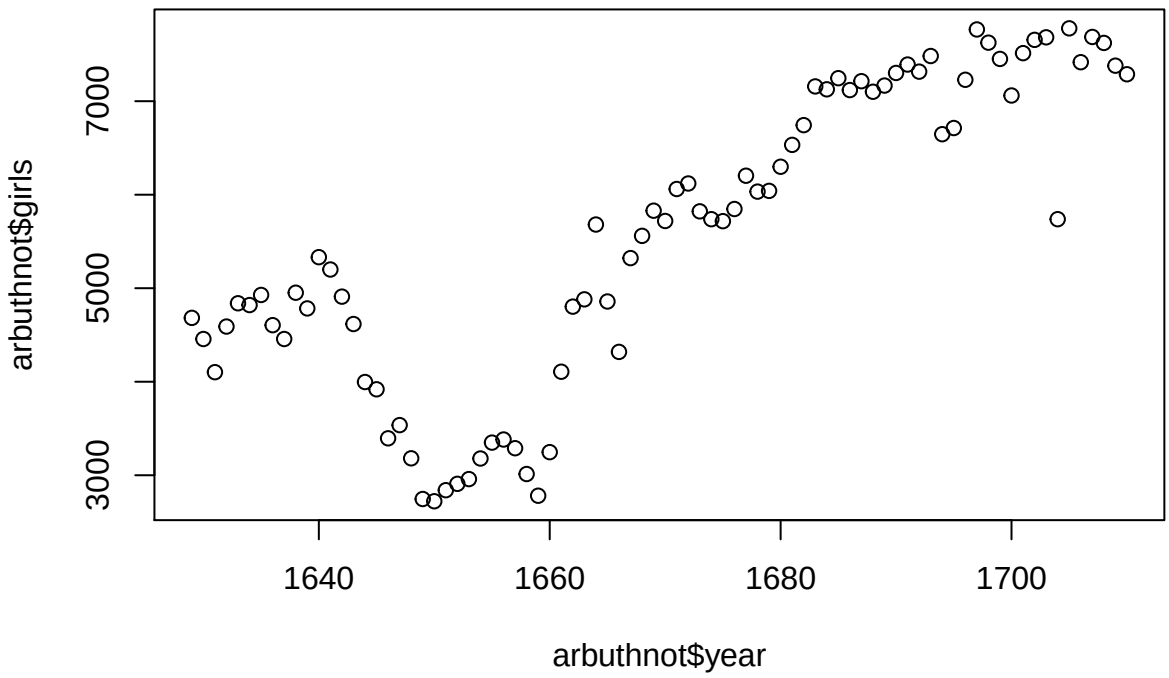
Exercise 1: What command would you use to extract just the counts of girls baptized?

You would use the command "arbuthnot$girls".

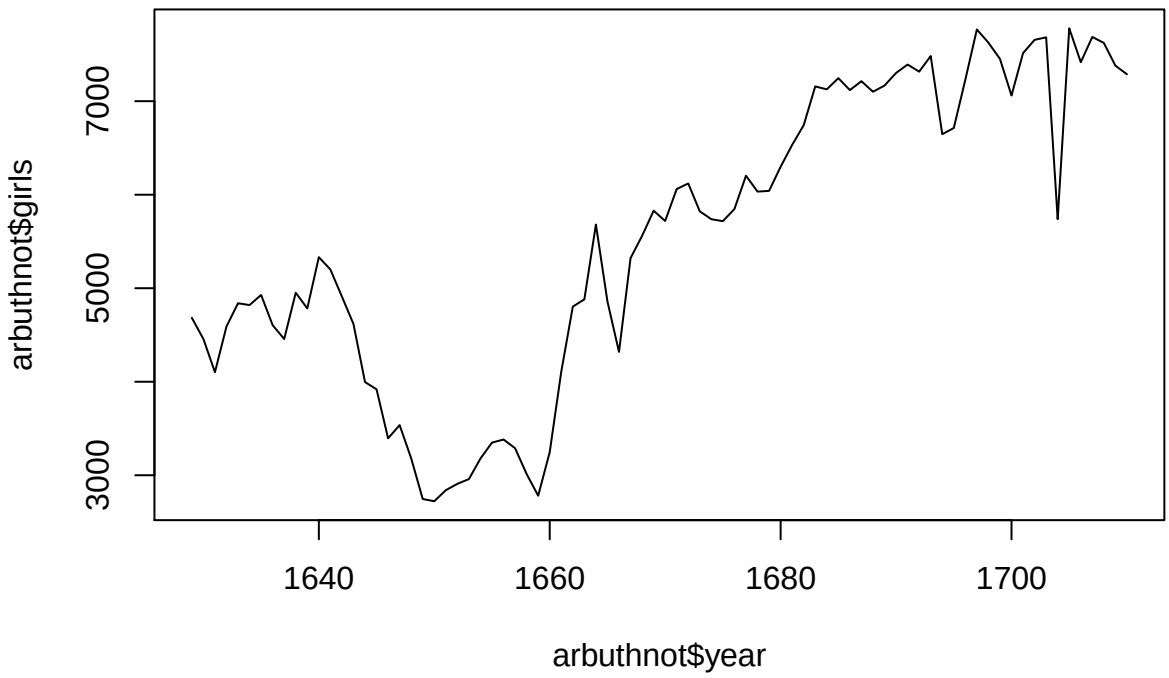
```
arbuthnot$girls
```

```
##  [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
```

```
plot(x = arbuthnot$year, y = arbuthnot$girls)
```




```
plot(x = arbuthnot$year, y = arbuthnot$girls, type = "l")
```




```
?plot
```

```
## starting httpd help server ... done
```

Exercise 2: Is there an apparent trend in the number of girls baptized over the years? How would you describe it?

Yes, there appears to be an overall upward trend in the number of girls baptized over the decades, although there was a big dip between 1640 and 1660.

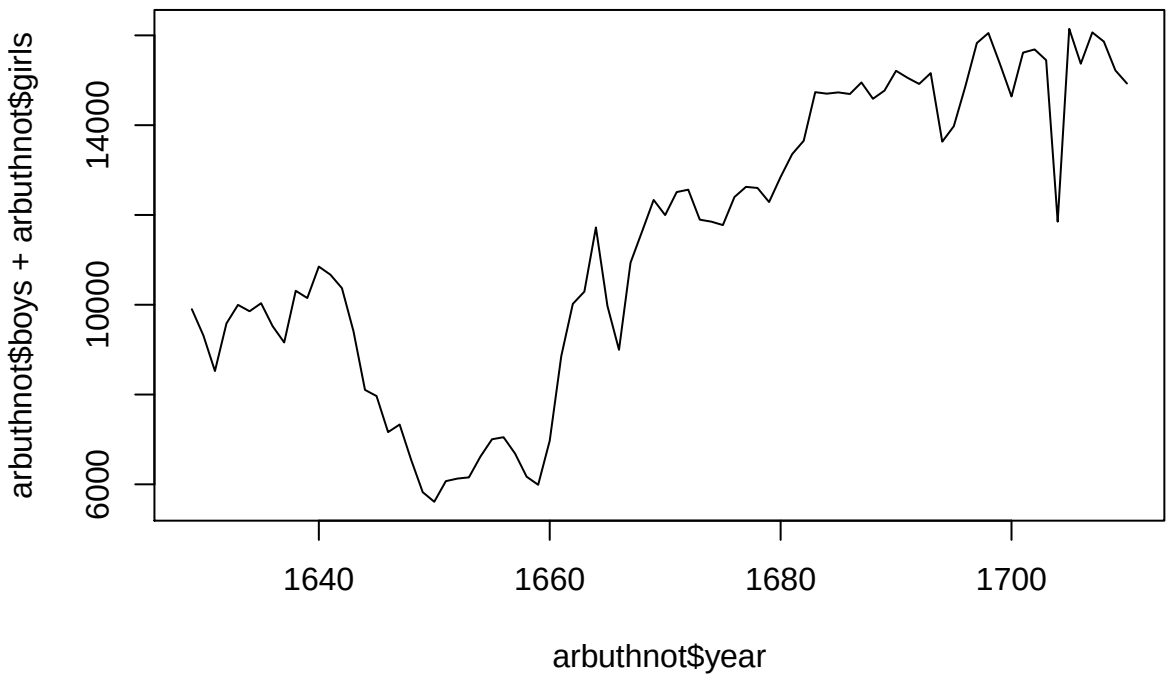
```
5218 + 4683
```

```
## [1] 9901
```

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

```
##  [1]  9901  9315  8524  9584  9997  9855 10034  9522  9160 10311 10150
## [12] 10850 10670 10370  9410  8104  7966  7163  7332  6544  5825  5612
## [23]  6071  6128  6155  6620  7004  7050  6685  6170  5990  6971  8855
## [34] 10019 10292 11722  9972  8997 10938 11633 12335 11997 12510 12563
## [45] 11895 11851 11775 12399 12626 12601 12288 12847 13355 13653 14735
## [56] 14702 14730 14694 14951 14588 14771 15211 15054 14918 15159 13632
## [67] 13976 14861 15829 16052 15363 14639 15616 15687 15448 11851 16145
## [78] 15369 16066 15862 15220 14928
```

```r
plot(arbuthnot$year, arbuthnot$boys + arbuthnot$girls, type = "l")
```




```r
5218 / 4683
```

```
## [1] 1.114243
```

```r
arbuthnot$boys / arbuthnot$girls
```

```
##  [1] 1.114243 1.089971 1.078011 1.088017 1.065923 1.044606 1.036120
##  [8] 1.067752 1.055194 1.082189 1.121656 1.034884 1.051923 1.112016
## [15] 1.038120 1.027521 1.032661 1.109867 1.073529 1.057215 1.121267
## [22] 1.061719 1.137676 1.107290 1.080095 1.082416 1.091371 1.084565
## [29] 1.032533 1.047793 1.153901 1.146905 1.156075 1.085988 1.108584
## [36] 1.063369 1.052697 1.083121 1.055242 1.092266 1.116143 1.097744
## [43] 1.064016 1.052778 1.043112 1.065354 1.059647 1.120575 1.035467
## [50] 1.088679 1.034100 1.039530 1.044237 1.024466 1.058536 1.062860
## [57] 1.032846 1.064054 1.072498 1.054359 1.060974 1.083128 1.036526
## [64] 1.039092 1.025792 1.050850 1.081931 1.055748 1.037981 1.104904
## [71] 1.061594 1.073219 1.078254 1.048981 1.010673 1.065354 1.075460
## [78] 1.072132 1.090022 1.080808 1.062331 1.048299
```

```r
5218 / (5218 + 4683)
```

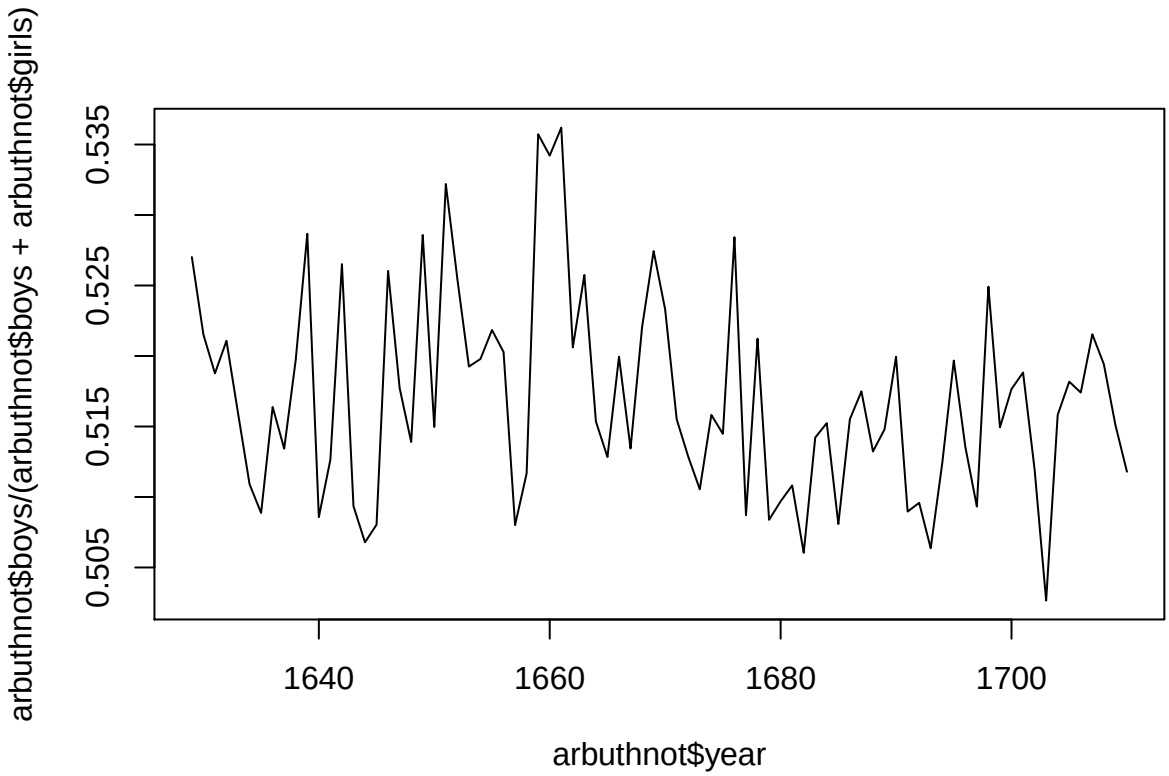
```
## [1] 0.5270175
```

```r
arbuthnot$boys / (arbuthnot$boys + arbuthnot$girls)
```

```
##  [1] 0.5270175 0.5215244 0.5187705 0.5210768 0.5159548 0.5109082 0.5088698
##  [8] 0.5163831 0.5134279 0.5197362 0.5286700 0.5085714 0.5126523 0.5265188
```

```
## [15] 0.5093518 0.5067868 0.5080341 0.5260366 0.5177305 0.5139059 0.5285837
## [22] 0.5149679 0.5322023 0.5254569 0.5192526 0.5197885 0.5218447 0.5202837
## [29] 0.5080030 0.5116694 0.5357262 0.5342132 0.5361942 0.5206108 0.5257482
## [36] 0.5153557 0.5128359 0.5199511 0.5134394 0.5220493 0.5274422 0.5232975
## [43] 0.5155076 0.5128552 0.5105507 0.5158214 0.5144798 0.5284297 0.5087122
## [50] 0.5212285 0.5083822 0.5096910 0.5108199 0.5060426 0.5142178 0.5152360
## [57] 0.5080788 0.5155165 0.5174905 0.5132301 0.5147925 0.5199527 0.5089677
## [64] 0.5095857 0.5063659 0.5123973 0.5196766 0.5135590 0.5093183 0.5249190
## [71] 0.5149385 0.5176583 0.5188268 0.5119526 0.5026541 0.5158214 0.5181790
## [78] 0.5174052 0.5215362 0.5194175 0.5151117 0.5117899
```

```
plot(arbuthnot$year, arbuthnot$boys / (arbuthnot$boys + arbuthnot$girls), type = "l")
```




```
arbuthnot$boys > arbuthnot$girls
```

```
##  [1] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [15] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [29] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [43] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [57] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [71] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
```

On Your Own

```
present <- read.csv("C:/Users/Kavya/Desktop/Education/MS Data Science/DATA 606 (Statistics and Probabil

head(present)
```

```
##   year    boys   girls
## 1 1940 1211684 1148715
## 2 1941 1289734 1223693
## 3 1942 1444365 1364631
## 4 1943 1508959 1427901
## 5 1944 1435301 1359499
## 6 1945 1404587 1330869
```

```
summary(present)
```

```
##       year           boys             girls
##  Min.   :1940   Min.   :1211684   Min.   :1148715
##  1st Qu.:1956   1st Qu.:1799857   1st Qu.:1711405
##  Median :1971   Median :1924868   Median :1831679
##  Mean   :1971   Mean   :1885600   Mean   :1793915
##  3rd Qu.:1986   3rd Qu.:2058524   3rd Qu.:1965538
##  Max.   :2002   Max.   :2186274   Max.   :2082052
```

---

1. What years are included in this dataset? What are the dimensions of the data frame and what are the variable or column names?

```
str(present$year)
```

```
##  int [1:63] 1940 1941 1942 1943 1944 1945 1946 1947 1948 1949 ...
```

```
dim(present)
```

```
## [1] 63  3
```

```
names(present)
```

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

This dataset includes all years from 1940 to 2002. The dimensions of this dataset are 63 rows by 3 columns. The variables included are "year", "boys", and "girls".

---

2. How do these counts compare to Arbuthnot's? Are they on a similar scale?

```
summary(arbuthnot$boys)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2890    4759    6073    5907    7576    8426
```

```
summary(arbuthnot$girls)
```

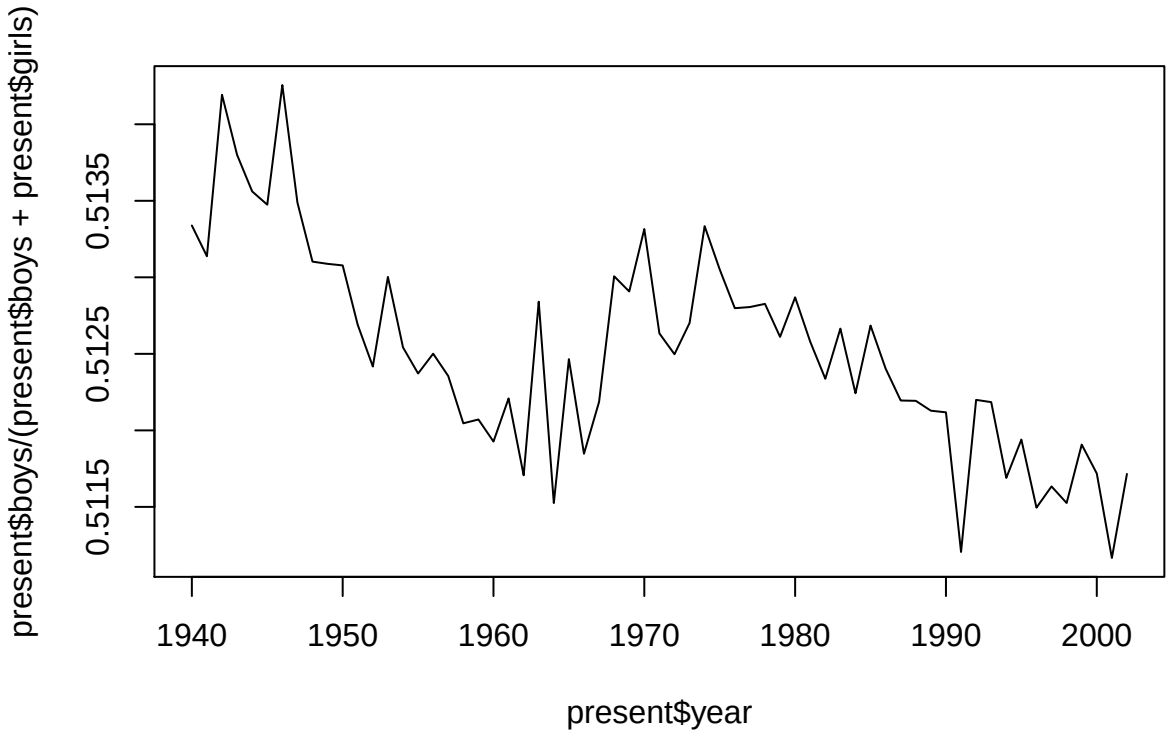
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2722    4457    5718    5535    7150    7779
```

The numbers are not comparable. The present dataset contains an average of 1.9 million male and 1.8 million female births per year. However, Arbuthnot contains an average of 5,900 male and 5,500 female births per year.

---

3. Make a plot that displays the boy-to-girl ratio for every year in the data set. What do you see? Does Arbuthnot's observation about boys being born in greater proportion than girls hold up in the U.S.? Include the plot in your response.

```
plot(present$year, present$boys / (present$boys + present$girls), type = "l")
```



Yes, Arbuthnot's observation appears to hold up. The ratio of boys to all children born between 1940 and 2002 has been above 0.5 for all years, although the ratio has gone down over the decades.

---

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

```
present$sum <- present$boys + present$girls
sorted <- present[order(present$sum),]

tail(sorted)
```

```
##    year    boys   girls     sum
## 23 1962 2132466 2034896 4167362
## 19 1958 2152546 2051266 4203812
## 20 1959 2173638 2071158 4244796
## 18 1957 2179960 2074824 4254784
## 21 1960 2179708 2078142 4257850
## 22 1961 2186274 2082052 4268326
```

We saw the greatest number of births in 1961 – a total of 4.3 million boys and girls were born.
