Lab1_Present

Andrew Guo

9/9/2021

The Present Data

First we import the Present data:

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

We can then determine its dimensions and identify the names of the columns in the data frame `present`:

```r
dim(present)
```

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

```r
names(present)
```

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

Exploration

```r
attach(present)
boys
```

```
##  [1] 1211684 1289734 1444365 1508959 1435301 1404587 1691220 1899876 1813852
## [10] 1826352 1823555 1923020 1971262 2001798 2059068 2073719 2133588 2179960
## [19] 2152546 2173638 2179708 2186274 2132466 2101632 2060162 1927054 1845862
## [28] 1803388 1796326 1846572 1915378 1822910 1669927 1608326 1622114 1613135
## [37] 1624436 1705916 1709394 1791267 1852616 1860272 1885676 1865553 1879490
## [46] 1927983 1924868 1951153 2002424 2069490 2129495 2101518 2082097 2048861
## [55] 2022589 1996355 1990480 1985596 2016205 2026854 2076969 2057922 2057979
```

```r
girls
```

```
##  [1] 1148715 1223693 1364631 1427901 1359499 1330869 1597452 1800064 1721216
## [10] 1733177 1730594 1827830 1875724 1900322 1958294 1973576 2029502 2074824
## [19] 2051266 2071158 2078142 2082052 2034896 1996388 1967328 1833304 1760412
## [28] 1717571 1705238 1753634 1816008 1733060 1588484 1528639 1537844 1531063
## [37] 1543352 1620716 1623885 1703131 1759642 1768966 1794861 1773380 1789651
## [46] 1832578 1831679 1858241 1907086 1971468 2028717 2009389 1982917 1951379
## [55] 1930178 1903234 1901014 1895298 1925348 1932563 1981845 1968011 1963747
```

```r
year
```

```
##  [1] 1940 1941 1942 1943 1944 1945 1946 1947 1948 1949 1950 1951 1952 1953 1954
## [16] 1955 1956 1957 1958 1959 1960 1961 1962 1963 1964 1965 1966 1967 1968 1969
## [31] 1970 1971 1972 1973 1974 1975 1976 1977 1978 1979 1980 1981 1982 1983 1984
## [46] 1985 1986 1987 1988 1989 1990 1991 1992 1993 1994 1995 1996 1997 1998 1999
## [61] 2000 2001 2002
```
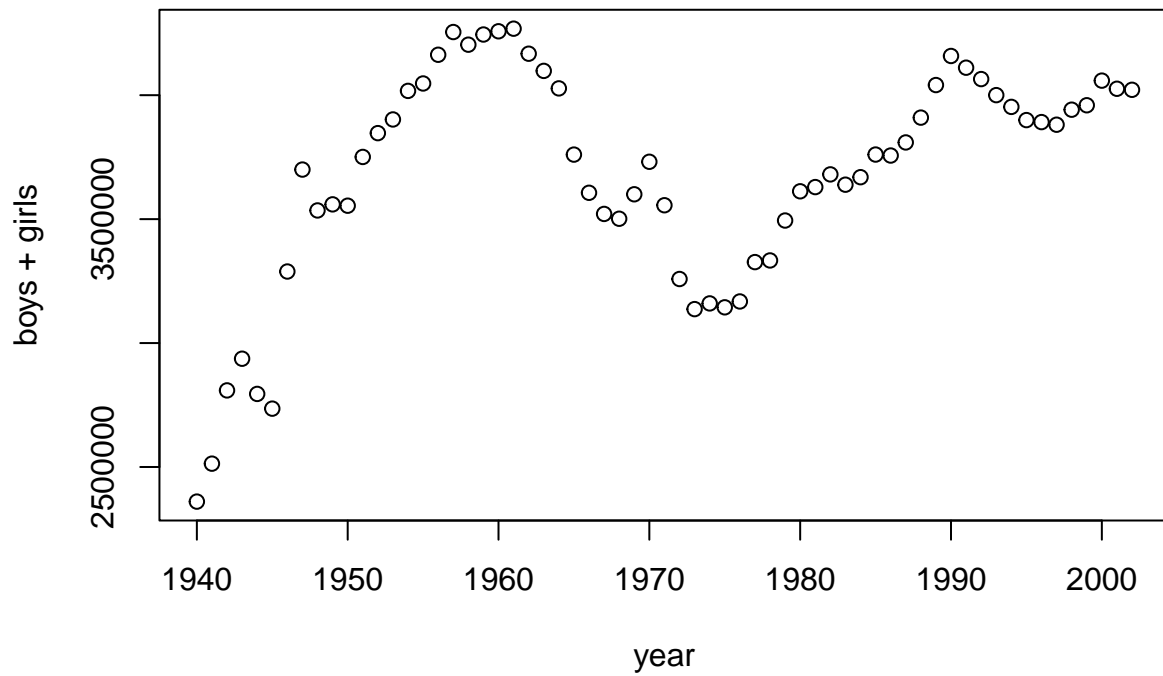
Plots
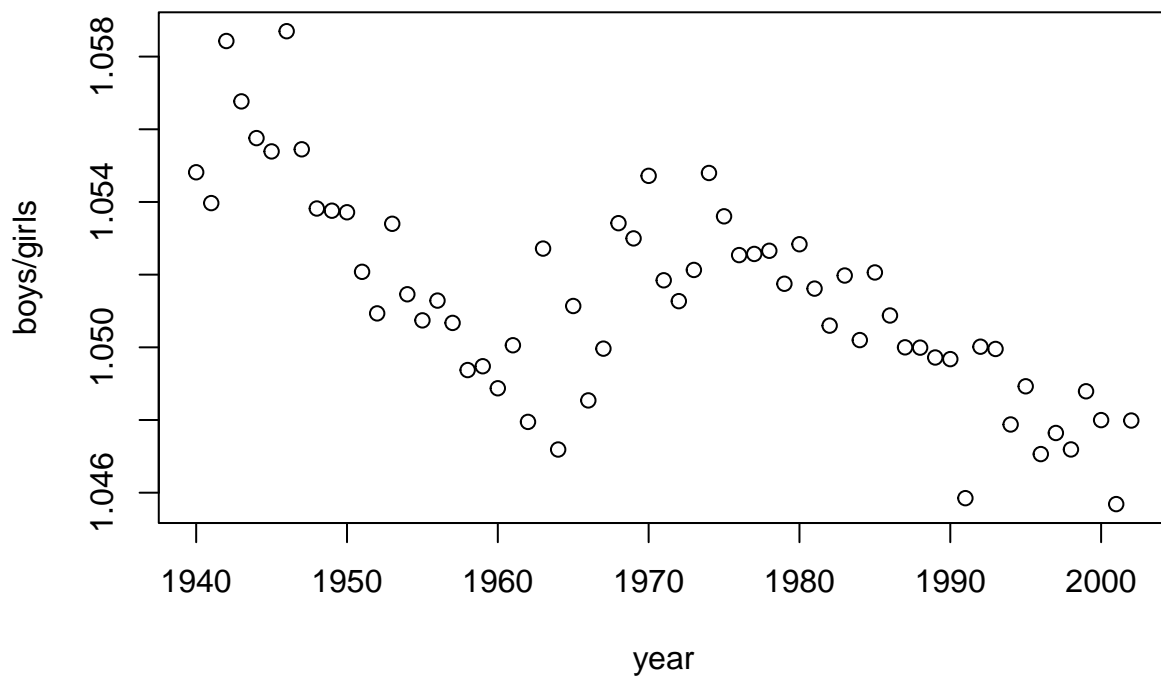
```
plot(year, boys, type = "p")
```
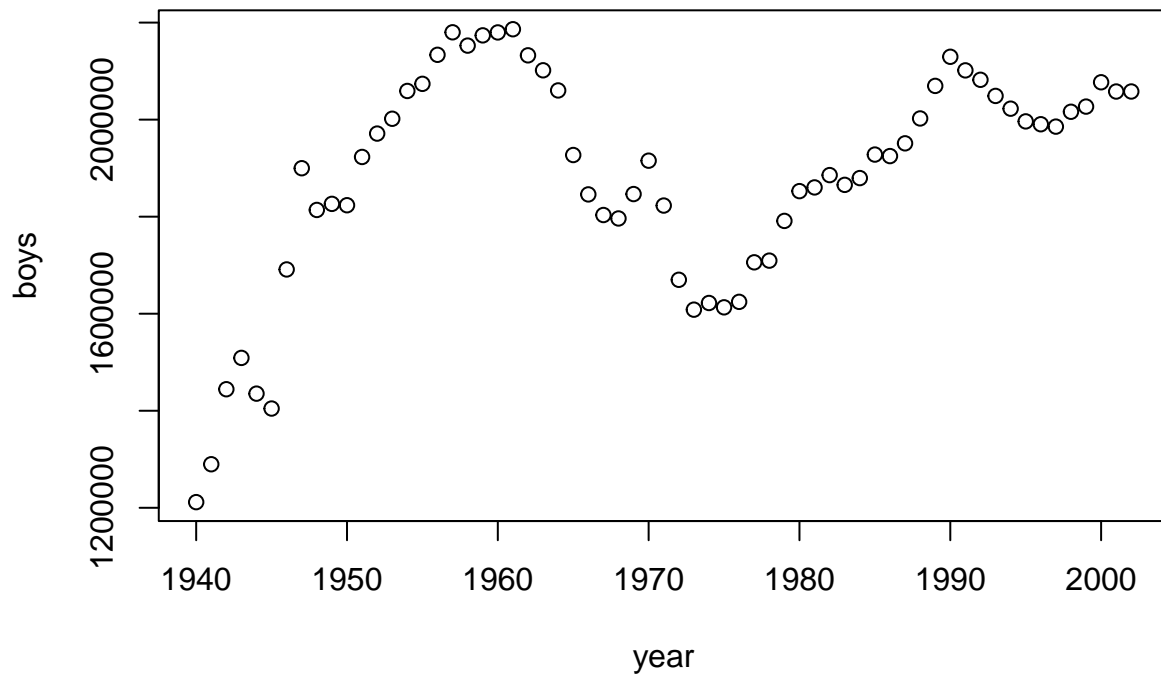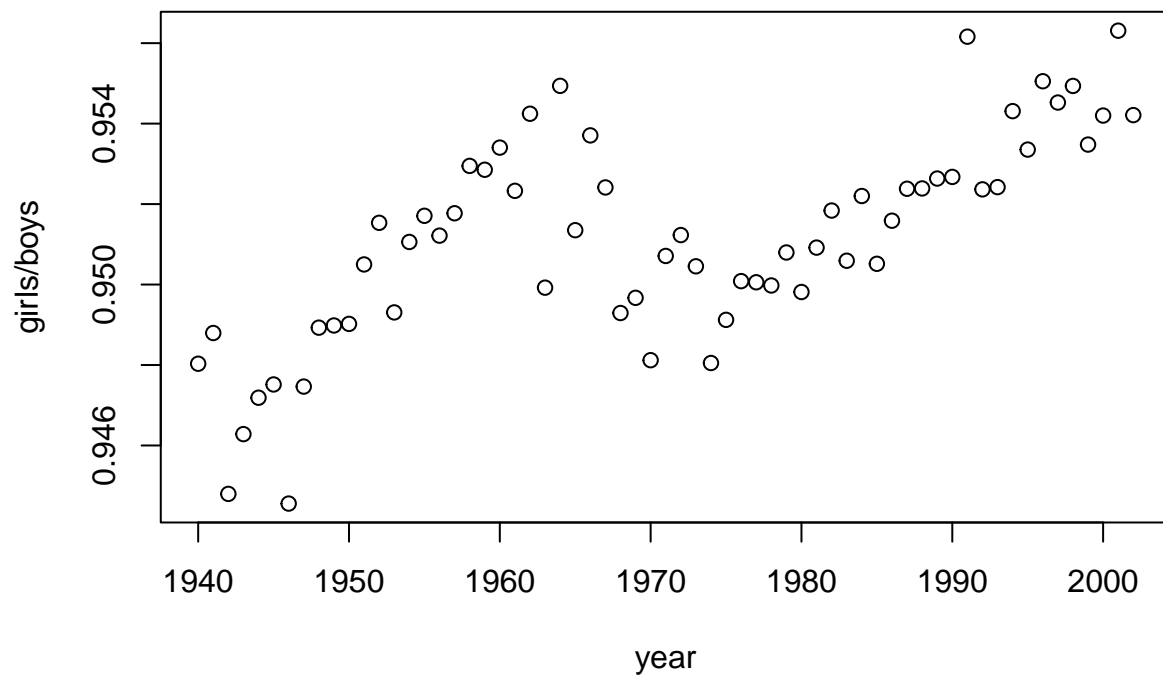


```
plot(year, girls, type = "p")
```
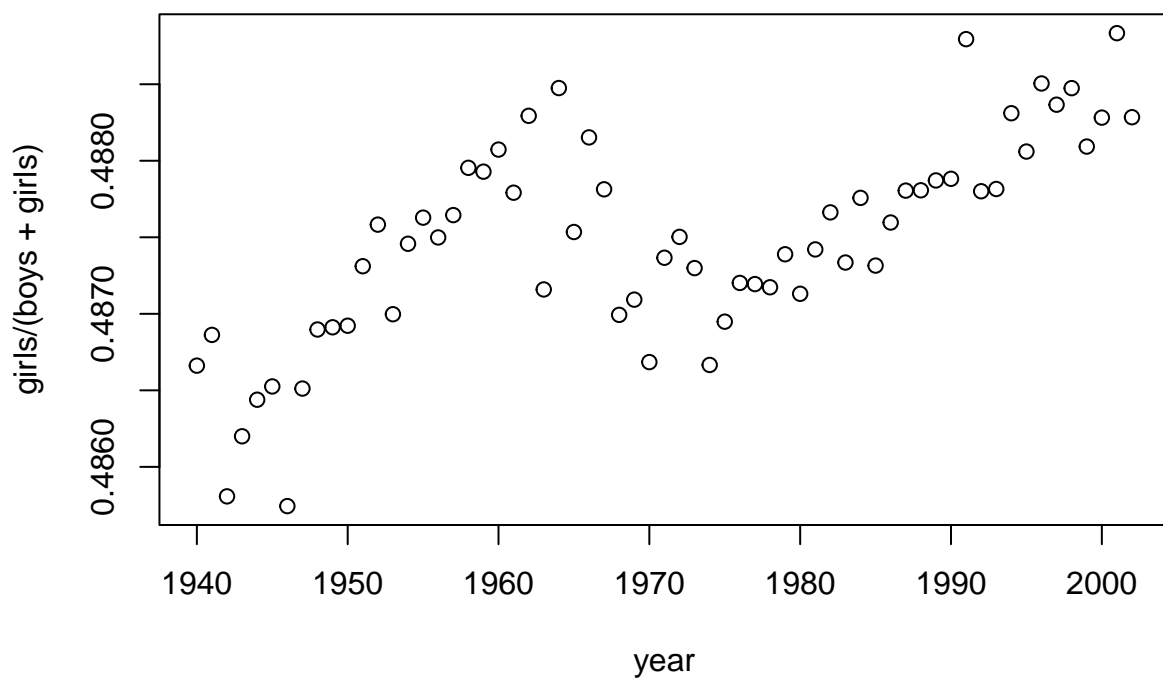


```
plot(year, boys+girls, type = "p")
```
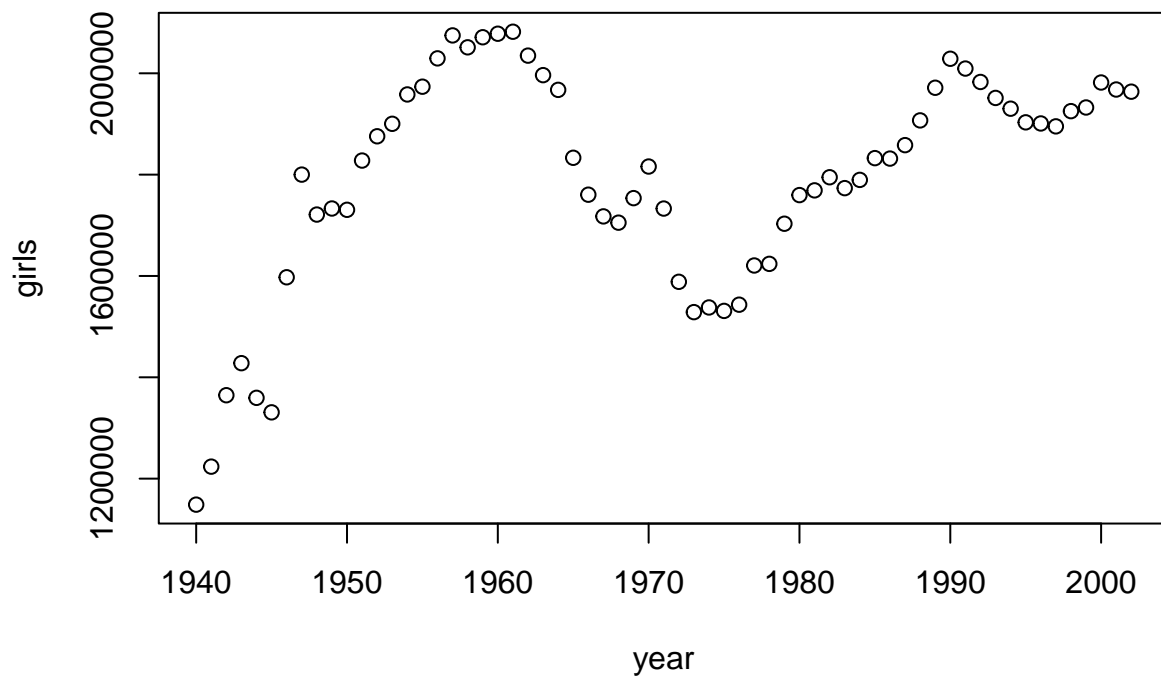
```
plot(year, boys/girls, type = "p")
```
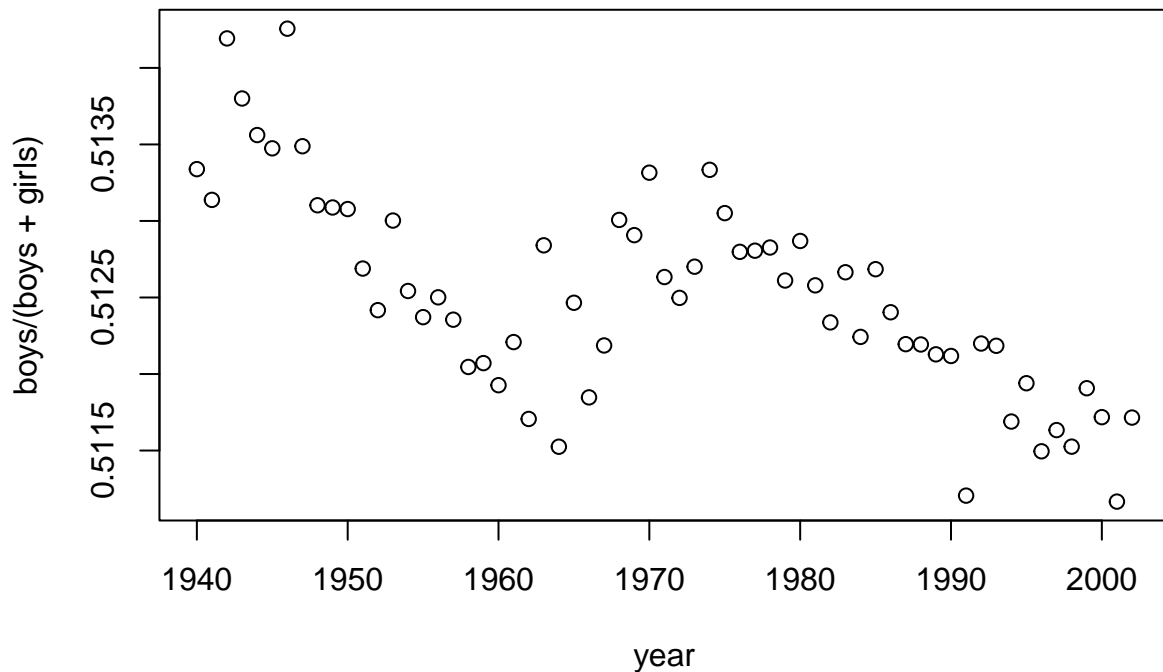


```
plot (year, girls/boys, type = "p")
```

3

```
plot(year, girls/(boys+girls), type = "p")
```



```
plot(year, boys/(boys+girls), type = "p")
```

4

Questions

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

To find the years included in this data set, we can use the attach() command:

```
attach(present)
```

```
## The following objects are masked from present (pos = 3):
##
##     boys, girls, year
year
```

```
##  [1] 1940 1941 1942 1943 1944 1945 1946 1947 1948 1949 1950 1951 1952 1953 1954
## [16] 1955 1956 1957 1958 1959 1960 1961 1962 1963 1964 1965 1966 1967 1968 1969
## [31] 1970 1971 1972 1973 1974 1975 1976 1977 1978 1979 1980 1981 1982 1983 1984
## [46] 1985 1986 1987 1988 1989 1990 1991 1992 1993 1994 1995 1996 1997 1998 1999
## [61] 2000 2001 2002
```

We find that the years in this data set are from 1940 to 2002.

To find the dimensions of the data frame, we can use the dim() command:

```
dim(present)
```

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

We find that there are a total of 63 rows and 3 columns.

To find the column names, we can use the names() command:

```
names(present)
```

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

We find that the column names are "year", "boys", and "girls".

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

To compare these counts with Arbuthnot's data set, we should first import that data set. To do so, we can use the source() command:

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

Afterwards, we will then use the dim(), names(), and attach() commands on the Arbuthnot data set to compare the properties of the different data sets' counts with one another.

```
dim(arbuthnot)
```

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

```
names(arbuthnot)
```

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

```
attach(arbuthnot)
```

```
## The following objects are masked from present (pos = 3):
##
##      boys, girls, year
```

```
## The following objects are masked from present (pos = 4):
##
##      boys, girls, year
```

```
year
```

```
##  [1] 1629 1630 1631 1632 1633 1634 1635 1636 1637 1638 1639 1640 1641 1642 1643
## [16] 1644 1645 1646 1647 1648 1649 1650 1651 1652 1653 1654 1655 1656 1657 1658
## [31] 1659 1660 1661 1662 1663 1664 1665 1666 1667 1668 1669 1670 1671 1672 1673
## [46] 1674 1675 1676 1677 1678 1679 1680 1681 1682 1683 1684 1685 1686 1687 1688
## [61] 1689 1690 1691 1692 1693 1694 1695 1696 1697 1698 1699 1700 1701 1702 1703
## [76] 1704 1705 1706 1707 1708 1709 1710
```

```
boys
```

```
##  [1] 5218 4858 4422 4994 5158 5035 5106 4917 4703 5359 5366 5518 5470 5460 4793
## [16] 4107 4047 3768 3796 3363 3079 2890 3231 3220 3196 3441 3655 3668 3396 3157
## [31] 3209 3724 4748 5216 5411 6041 5114 4678 5616 6073 6506 6278 6449 6443 6073
## [46] 6113 6058 6552 6423 6568 6247 6548 6822 6909 7577 7575 7484 7575 7737 7487
## [61] 7604 7909 7662 7602 7676 6985 7263 7632 8062 8426 7911 7578 8102 8031 7765
## [76] 6113 8366 7952 8379 8239 7840 7640
```

```
girls
```

```
##  [1] 4683 4457 4102 4590 4839 4820 4928 4605 4457 4952 4784 5332 5200 4910 4617
## [16] 3997 3919 3395 3536 3181 2746 2722 2840 2908 2959 3179 3349 3382 3289 3013
## [31] 2781 3247 4107 4803 4881 5681 4858 4319 5322 5560 5829 5719 6061 6120 5822
## [46] 5738 5717 5847 6203 6033 6041 6299 6533 6744 7158 7127 7246 7119 7214 7101
## [61] 7167 7302 7392 7316 7483 6647 6713 7229 7767 7626 7452 7061 7514 7656 7683
## [76] 5738 7779 7417 7687 7623 7380 7288
```

Based on these computations, we observe several features.

The dim() command reveals differences in the number of observations between the two data sets. While the present data set had a total of 63 rows, there are 82 rows in the arbuthnot data set.

The names() command reveals that the names of the rows in present and arbuthnot are the same; all have variable names called "year", "boys", and "girls" that correspond respectively to the year of the observation,

the number of boys in a given year, and the number of girls in a given year.

The attach() command also reveals differences in the years of data collection. While the present data set accounts for years 1940-2002, the arbuthnot data set accounts for years 1629-1710. The present data set covers approximately 62 years of data, while the arbuthnot data set covers approximately 81 years. Additionally, the recorded number of boys and girls in the present data set is substantially larger than the recorded number of boys and girls in the arbuthnot data set. This can be attributed to overall population growth since the seventeenth/eighteenth century and that population numbers can be counted through census data rather than baptisms which can only account for religious families.

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.

```
attach(arbuthnot)
```

```
## The following objects are masked from arbuthnot (pos = 3):
##
##      boys, girls, year
```

```
## The following objects are masked from present (pos = 4):
##
##      boys, girls, year
```

```
## The following objects are masked from present (pos = 5):
##
##      boys, girls, year
```
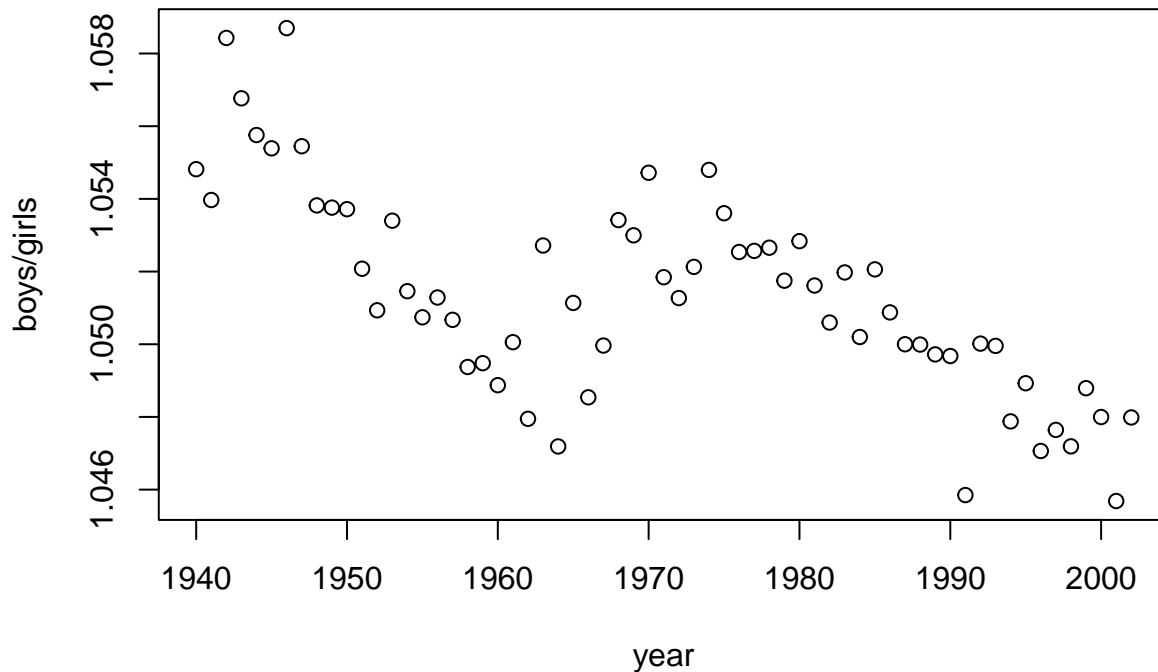
```
plot(year, boys/girls, type = "p")
```
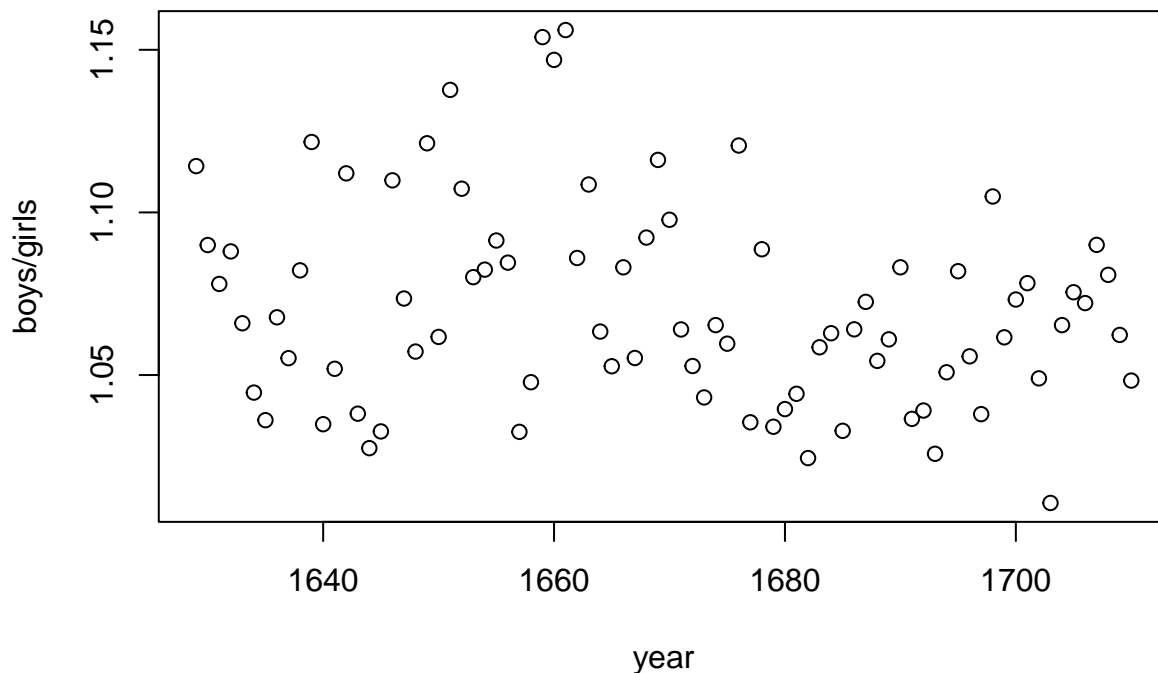


```
attach(present)
```

```
## The following objects are masked from arbuthnot (pos = 3):
##
##      boys, girls, year
```

```
## The following objects are masked from arbuthnot (pos = 4):
```

```
##
##     boys, girls, year

## The following objects are masked from present (pos = 5):
##
##     boys, girls, year

## The following objects are masked from present (pos = 6):
##
##     boys, girls, year
```

```
plot(year, boys/girls, type = "p")
```


Arbuthnot's observation was that on average, there were more boys than girls as the ratio of boys to girls was always positive. This still holds true in the present data set, as the ratio never dips below one. However, the spread of the present data set is much smaller, as it varies between around 1.06 and 1.046, whereeas the arbuthnot data showed a spread between around 1.15 and 1.0. Additionally, there was a clear, right-skewed shape in the present data set, showcasing that the 1940s had a higher proportion of boys to girls, but as time passed, this ratio decreased to a significant amount around the mid 1960s, moderately went back up around the early 1970s, but soon dipped back down as the 2000s drew closer. Artbuthnot's observation still stands, but as time has progressed, it has increasing been a much closer call.

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

attach(present)

```
attach(present)
```

```
## The following objects are masked from present (pos = 3):
##
##     boys, girls, year

## The following objects are masked from arbuthnot (pos = 4):
##
##     boys, girls, year

## The following objects are masked from arbuthnot (pos = 5):
```

```
##
##     boys, girls, year

## The following objects are masked from present (pos = 6):
##
##     boys, girls, year

## The following objects are masked from present (pos = 7):
##
##     boys, girls, year
```

```
year[which.max(boys+girls)]
```

```
## [1] 1961
```

```
plot(year, boys+girls, type = "p")
```
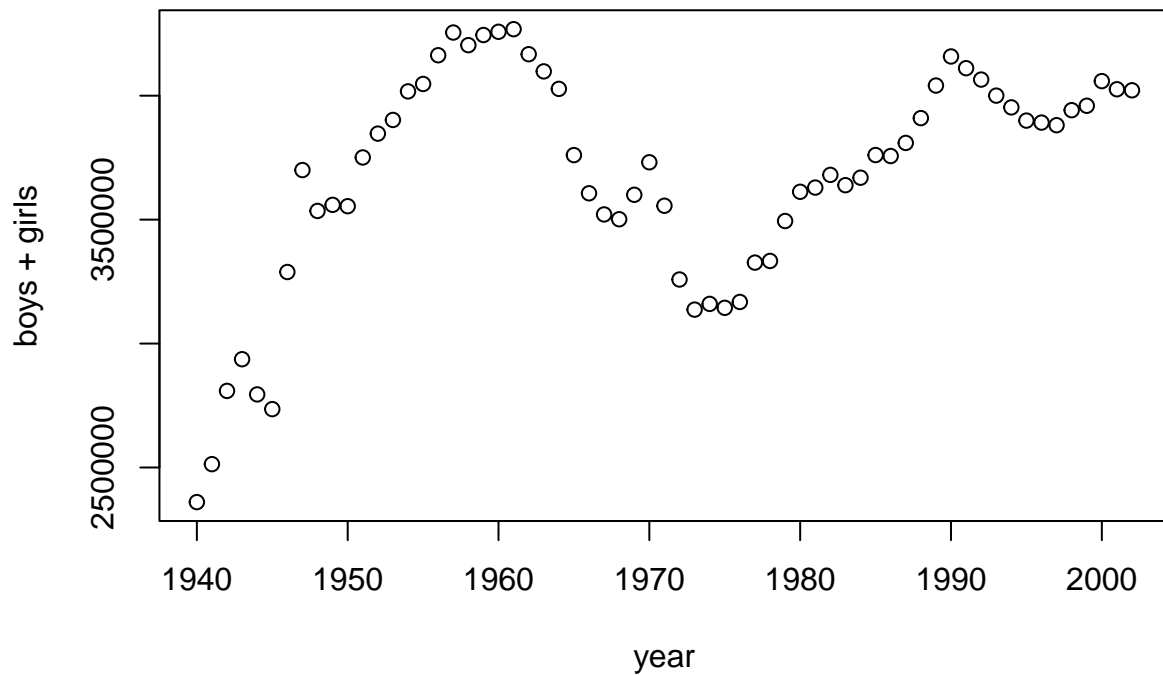

IF we run the which.max() command on the present data set, we find that the 22nd index holds the most boys and girls born in a year. If we take this index into year[] by using year[which.max(boys+girls)], then we confirm that this occurred at the year 1961. This observation is backed up by the graph of total births over the years in the present data set, which reveals its maximum point around the 1960/1961 mark.