Lab 1: Intro to R

Code

Julian Adames-Ng

2022-01-28

Hide

```
library(tidyverse)
library(openintro)
```

Exercise 1

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

The vector that was returned has 82 elements in it. These correnspond to the count of girls baptized.

Hide

```
arbuthnot$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
```

Hide

```
#5218 + 4683

#sum of boys column and girls column
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
```

Hide

```
#SCRAPWORK
#scatterplot and curve
#ggplot(data = arbuthnot, aes(x = year, y = total)) + geom_line()
#ggplot(data = arbuthnot, aes(x = year, y = total)) + geom_point()


#ratio of births single case
#5218 / 4683

#proportion of boys single case
#5218 / (5218 + 4683)
```

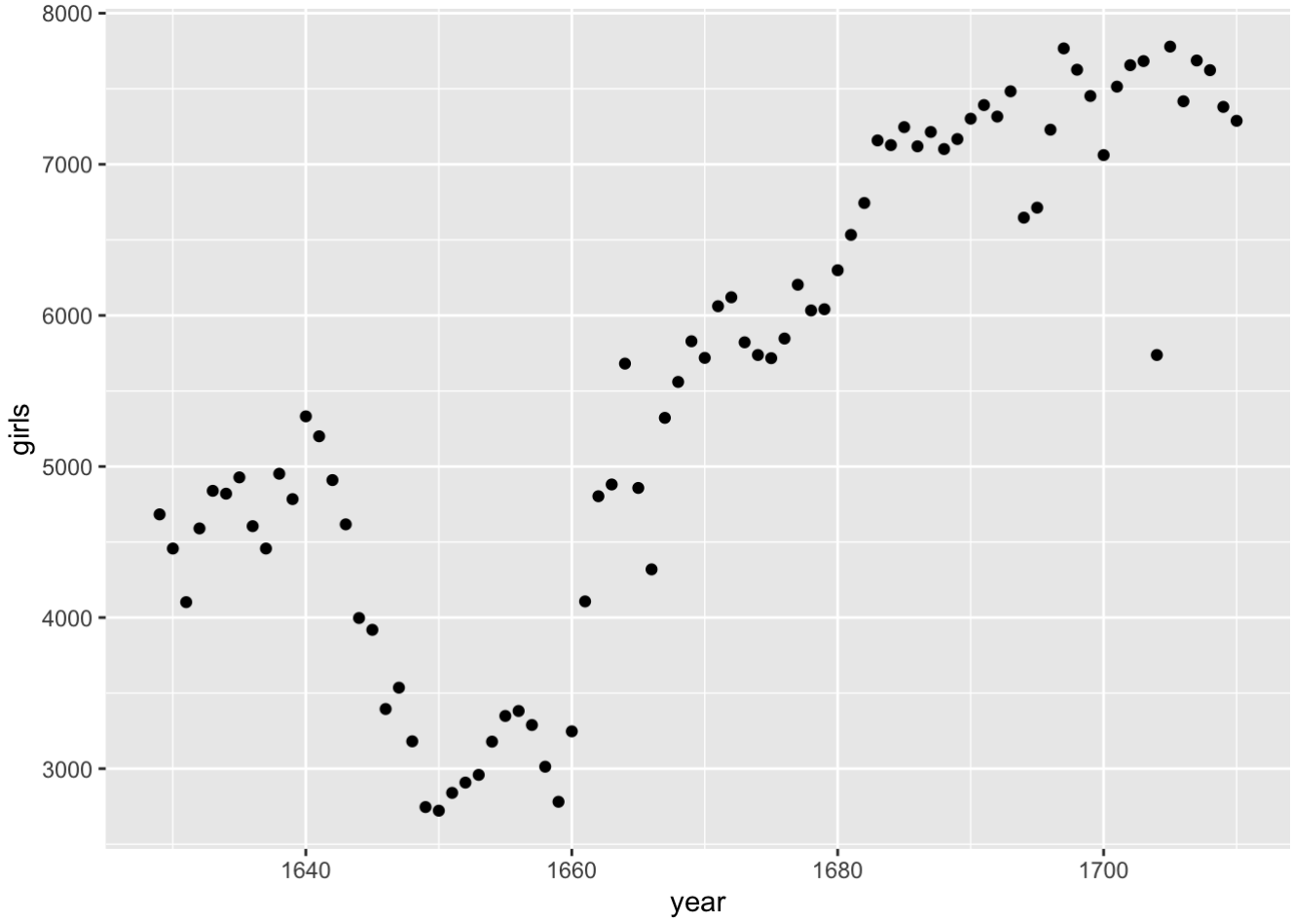
Exercise 2

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.)

There's an increasing trend in girl baptisms moving toward the year 1700. When compared to boys though, they both showed an increasing trend. A possibility is that this can be due to advancements that led up to industrialization like the printing press which may have allowed a faster spread of the Christianity.

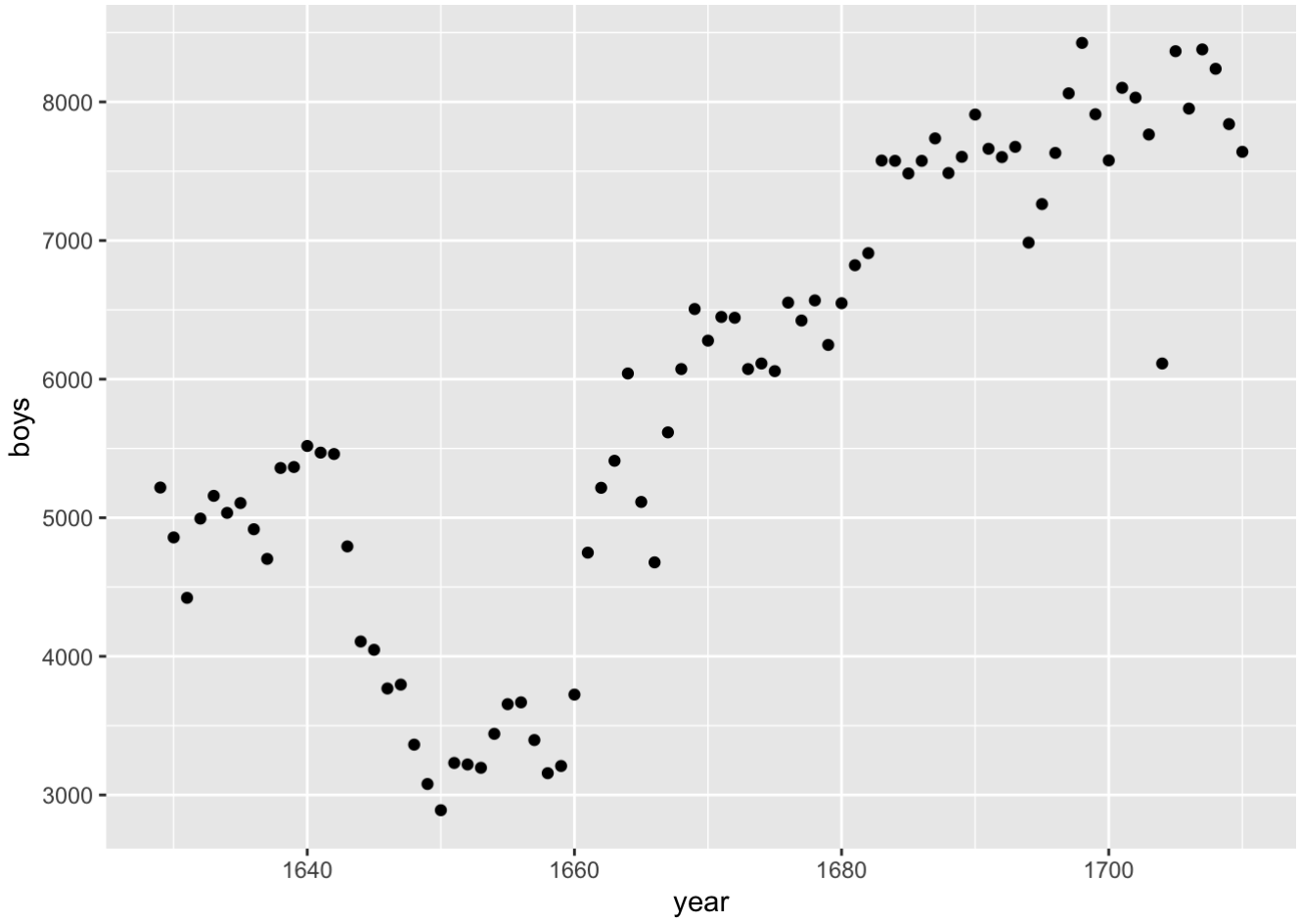
Hide

```
#compare girl births to boy births
ggplot(data = arbuthnot, aes(x = year, y = girls)) + geom_point()
```



Hide

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



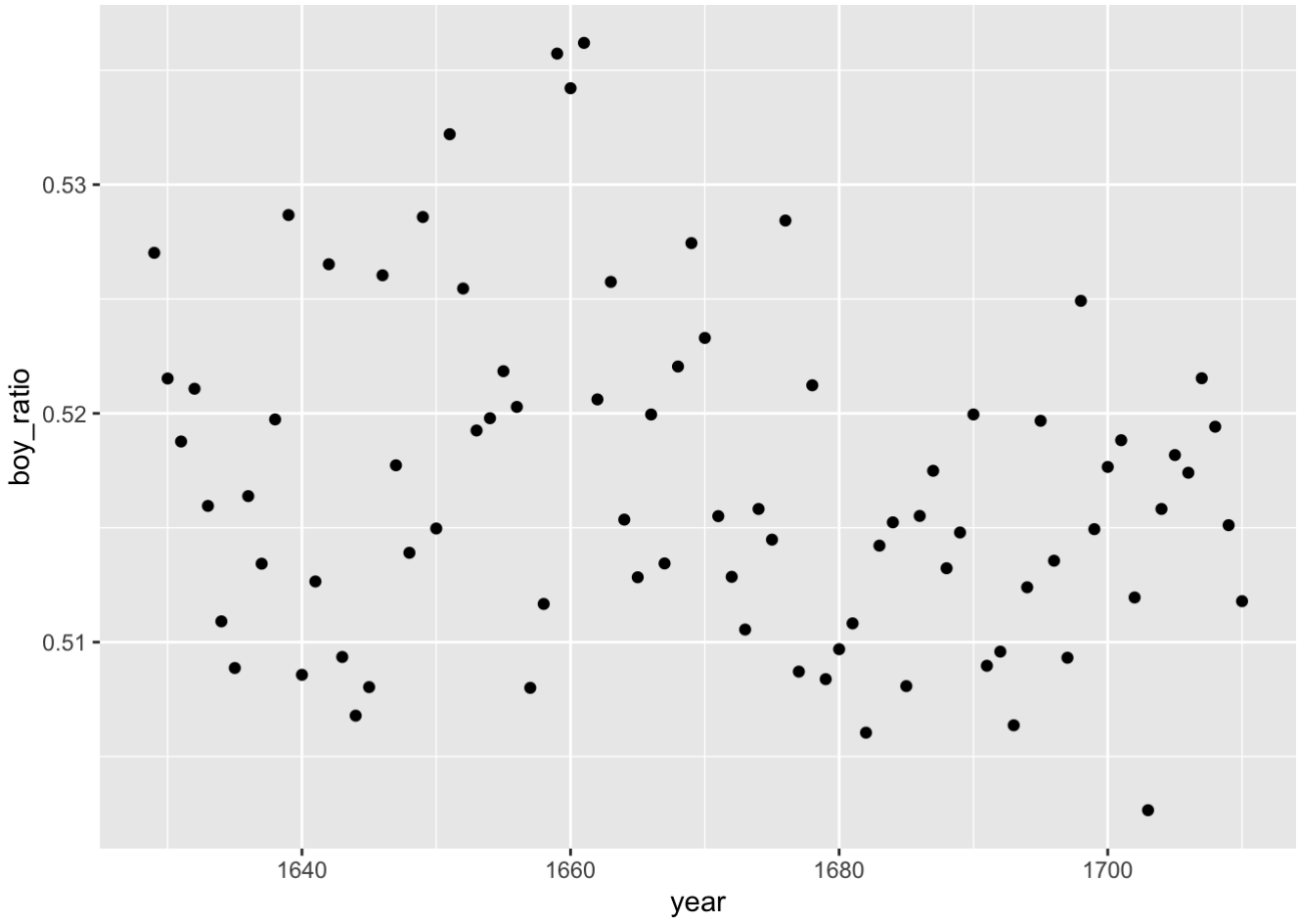
Exercise 3

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

The data points are scattered, but it seems that the proportion of boys over has stayed consistently above 0.50 for the entirety of Dr. Arbuthnot's study. The vector "more_boys" confirms this as each of its element values assesses whether there were more boys than girls during the given year and they all hold true.

Hide

```
# x %>% f(y) means f(x, y)
arbuthnot <- arbuthnot %>% mutate(total = boys + girls)

#ratio of births vector case
arbuthnot <- arbuthnot %>% mutate(boy_to_girl_ratio = boys / girls)
#arbuthnot$boy_to_girl_ratio

#proportion of boys vector case
arbuthnot <- arbuthnot %>% mutate(boy_ratio = boys / total)
ggplot(data = arbuthnot, aes(x = year, y = boy_ratio)) + geom_point()
```



Hide

```
#arbuthnot$boy_ratio

#proportion of girls vector case
arbuthnot <- arbuthnot %>% mutate (girl_ratio = girls / total)
#ggplot(data = arbuthnot, aes(x = year, y = girl_ratio)) + geom_line()


#ggplot(data = arbuthnot, aes(x = year, y = boy_to_girl_ratio)) + geom_point()

#ggplot(data = arbuthnot, aes(x = boys, y = girls)) + geom_point()
arbuthnot <- arbuthnot %>% mutate(more_boys = boys > girls)
arbuthnot$more_boys
```

```
##  [1] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [16] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [31] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [46] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [61] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [76] TRUE TRUE TRUE TRUE TRUE TRUE TRUE
```

Exercise 4

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

The data set includes years from 1940 to 2002.

The dimensions of the data frame are 63 rows by 3 columns where labeled as “year”, “boys”, and “girls” respectively.

Hide

```
data('present', package = 'openintro')

present
```

```
## # A tibble: 63 × 3
##     year    boys   girls
##    <dbl>   <dbl>   <dbl>
##  1  1940 1211684 1148715
##  2  1941 1289734 1223693
##  3  1942 1444365 1364631
##  4  1943 1508959 1427901
##  5  1944 1435301 1359499
##  6  1945 1404587 1330869
##  7  1946 1691220 1597452
##  8  1947 1899876 1800064
##  9  1948 1813852 1721216
## 10  1949 1826352 1733177
## # … with 53 more rows
```

Hide

```
present$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
```

Exercise 5

How do these counts compare to Arbuthnot’s? Are they of a similar magnitude?

The counts from the “present” data set are magnitudes larger than the counts from Dr. Arbuthnot’s original data set. There are less data points, but the population of boys and girls has increased over 200-fold.

Hide

```
data('present', package = 'openintro')


#compare totals
present %>% summarize(minP = min(boys), maxP = max(boys))
```

```
## # A tibble: 1 × 2
##      minP    maxP
##     <dbl>   <dbl>
## 1 1211684 2186274
```

Hide

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

```
## # A tibble: 1 × 2
##    minA  maxA
##   <int> <int>
## 1  2890  8426
```

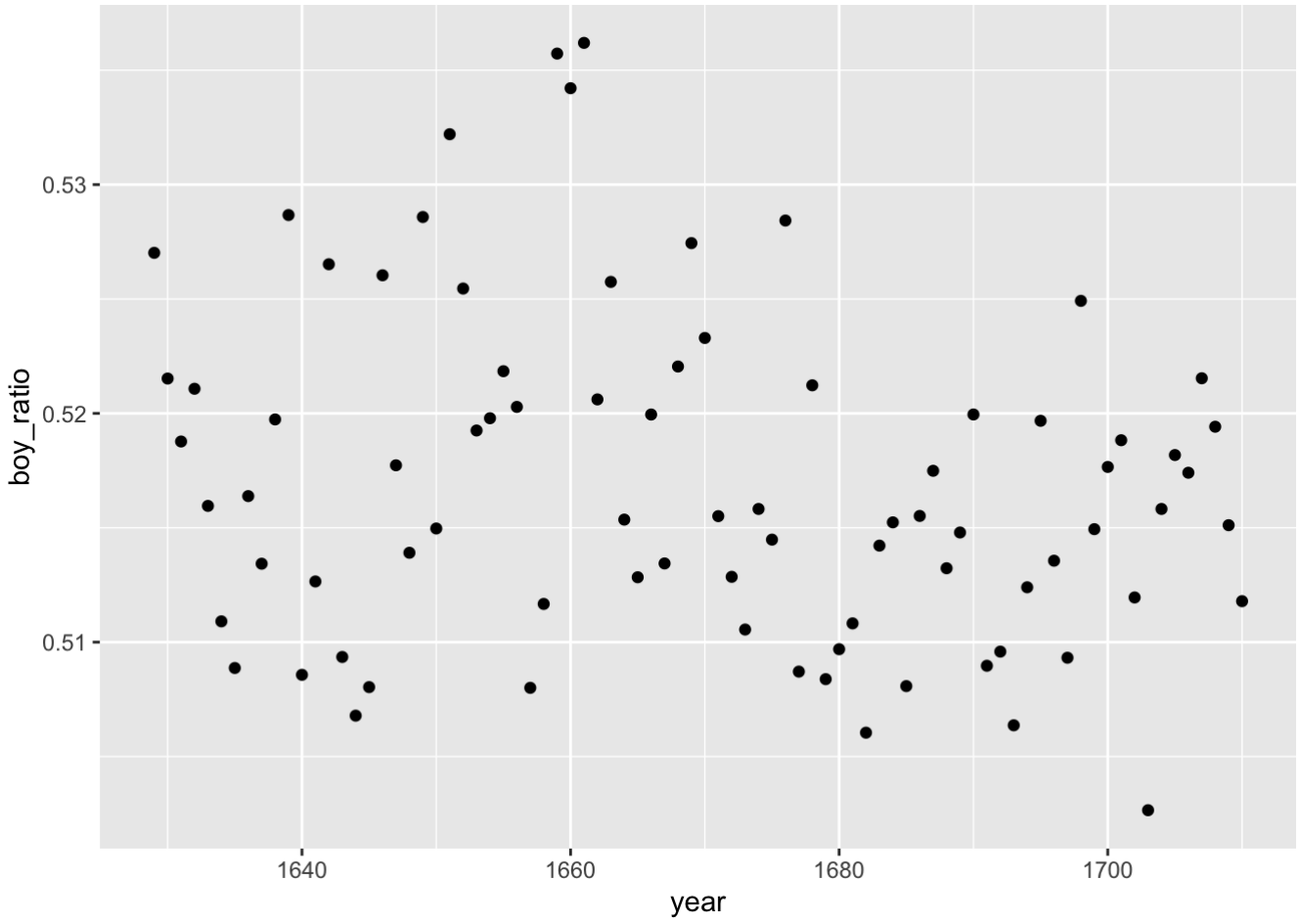
Exercise 6

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.? Include the plot in your response. Hint: You should be able to reuse your code from Exercise 3 above, just replace the dataframe name.

I plotted the Boy Proportion vs Year graphs for each data set. Although the proportion decreased for the entirety of the study for the "present" data set, it also remained above 0.50. The vector "more_boys" confirms this. Dr. Arbuthnot's findings hold true still.

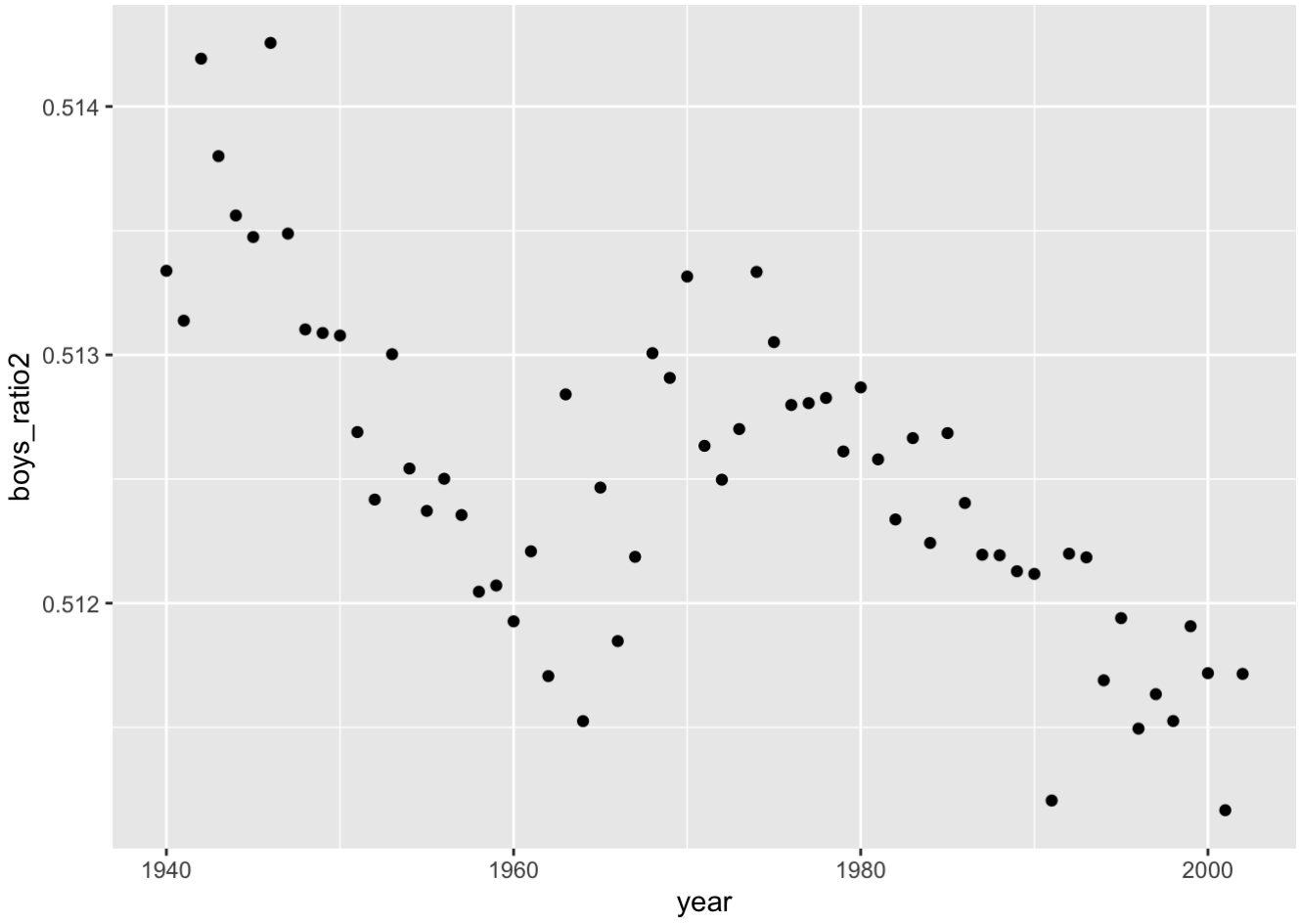
Hide

```
#create a new column for totals in present data set
present <- present %>% mutate(total = boys + girls)

#create a column for proportion of boys
present <- present %>% mutate(boys_ratio2 = boys / total )

#compare boy ratios
ggplot(data = arbuthnot, aes(x = year, y = boy_ratio)) + geom_point()
```



Hide

```
ggplot(data = present, aes(x = year, y = boys_ratio2)) + geom_point()
```



Hide

```
#truth column
present <- present %>% mutate(more_boys = boys > girls)
present$more_boys
```

```
##  [1] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [16] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [31] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [46] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [61] TRUE TRUE TRUE
```

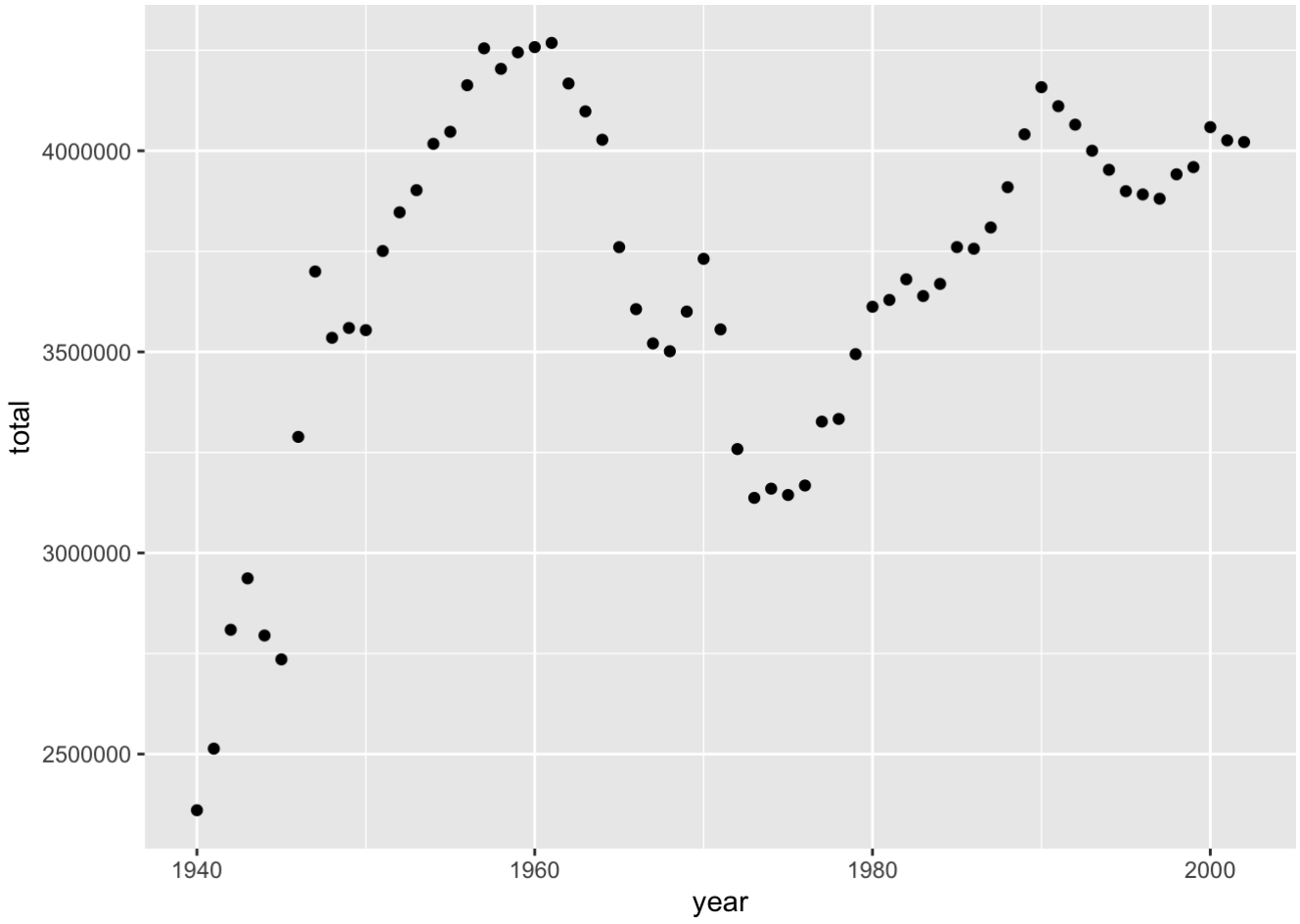
Exercise 7

In what year did we see the most total number of births in the U.S.? Hint: First calculate the totals and save it as a new variable. Then, sort your dataset in descending order based on the total column. You can do this interactively in the data viewer by clicking on the arrows next to the variable names. To include the sorted result in your report you will need to use two new functions: arrange (for sorting). We can arrange the data in a descending order with another function: desc (for descending order). The sample code is provided below.

The visual shows that the largest data point seems to correspond to a year in the early 1960s. Using the arrange() and desc() functions, we can confirm that the largest total came from the year 1961 with a total of 4,268,326 children.

Hide

```
#new totals column created in previous exercise

#visual check
ggplot(data = present, aes(x = year, y = total)) + geom_point()
```



Hide

```
#sort present data set by total in descending order
arrange(present, desc(total))
```

```
## # A tibble: 63 × 6
##     year    boys   girls   total boys_ratio2 more_boys
##    <dbl>   <dbl>   <dbl>   <dbl>       <dbl> <lgl>
##  1  1961 2186274 2082052 4268326       0.512 TRUE
##  2  1960 2179708 2078142 4257850       0.512 TRUE
##  3  1957 2179960 2074824 4254784       0.512 TRUE
##  4  1959 2173638 2071158 4244796       0.512 TRUE
##  5  1958 2152546 2051266 4203812       0.512 TRUE
##  6  1962 2132466 2034896 4167362       0.512 TRUE
##  7  1956 2133588 2029502 4163090       0.513 TRUE
##  8  1990 2129495 2028717 4158212       0.512 TRUE
##  9  1991 2101518 2009389 4110907       0.511 TRUE
## 10  1963 2101632 1996388 4098020       0.513 TRUE
## # … with 53 more rows
```