Data 606-Lab 1: Intro to R

Code

Sean Connin

2021-02-05

Hide

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

Exercise 1

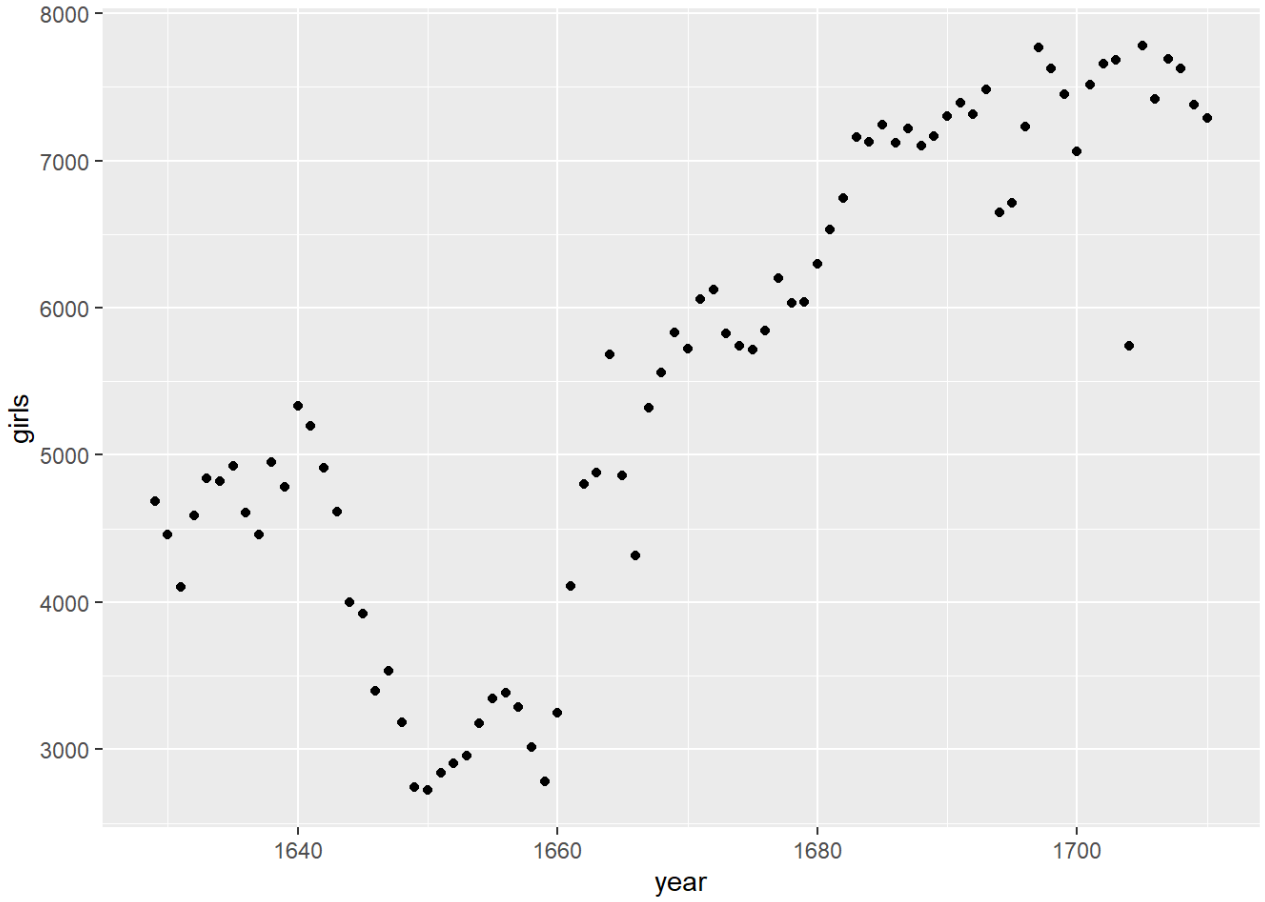
Hide

```
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

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



Exercise 2

The number of girls baptized each year declined during the period 1640 and 1660 (to < 3000) before increasing again to levels >7000 after 1680.

This pattern is apparent in the annual counts for boy baptisms as well.

Annual boys:girls counts vary (without obvious temporal pattern) between 1.04 and 1.16 over the same time-span. More boys are baptized each year than girls.

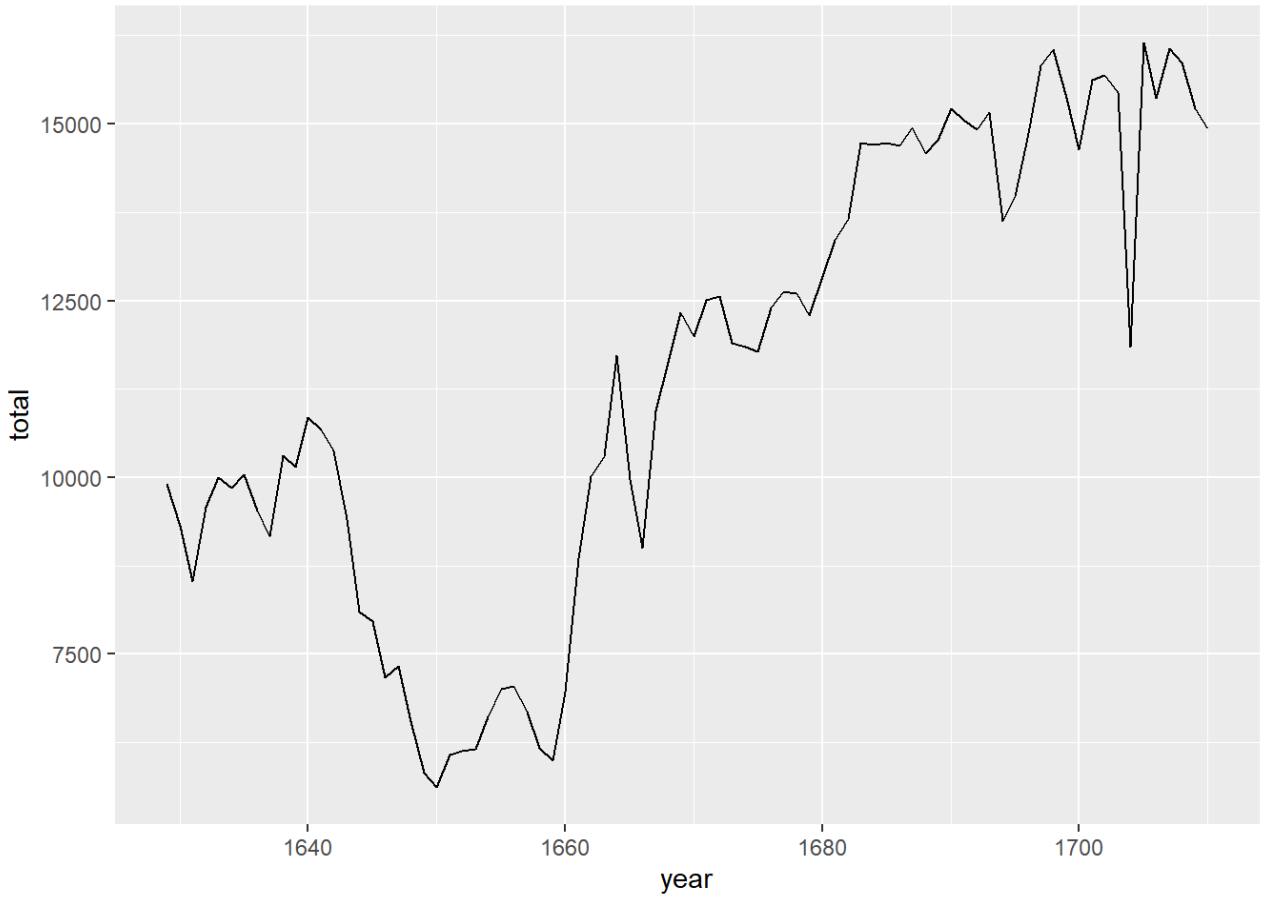
Hide

```
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
```

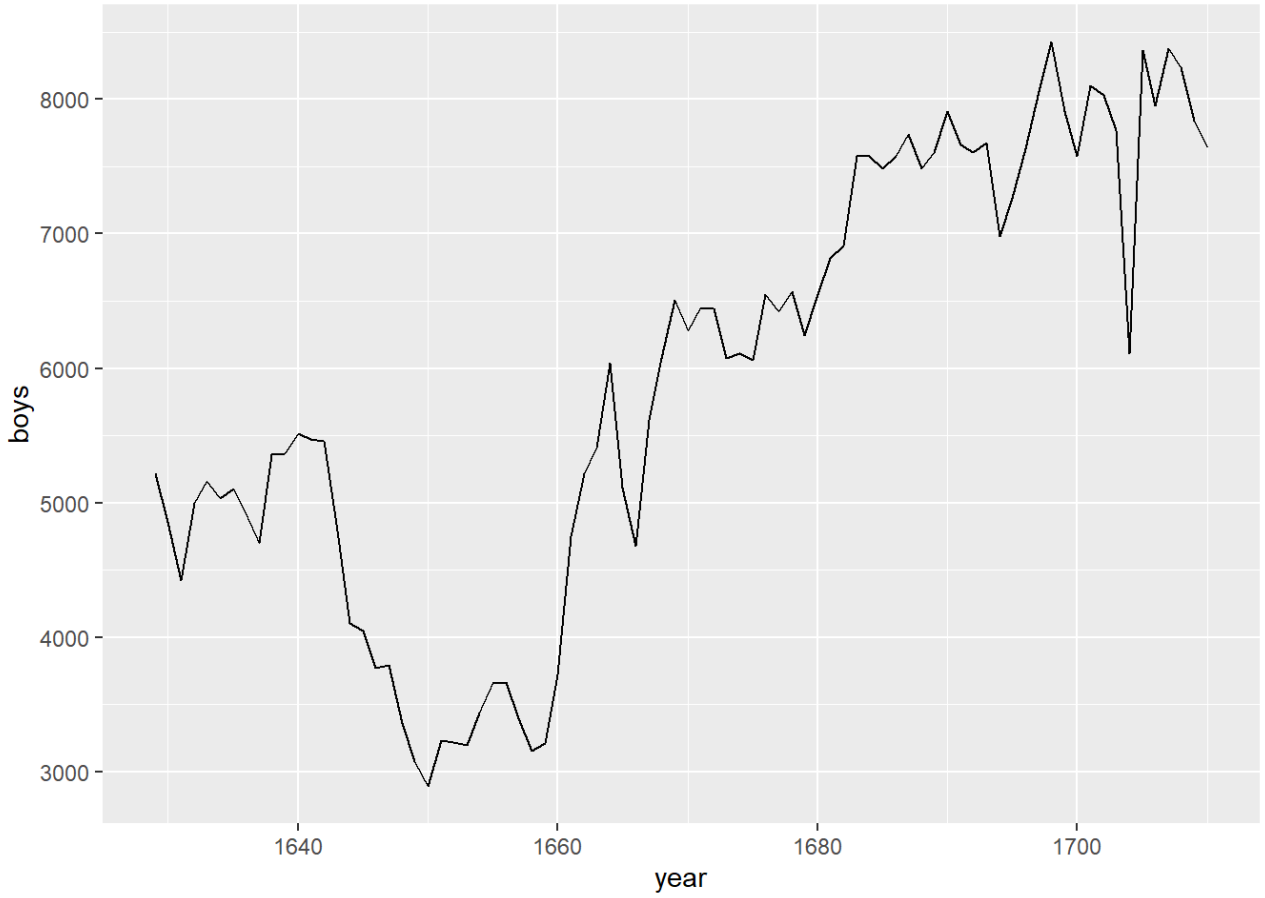
Hide

```
# create new column to sum boys + girls count by year

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

# create scatterplot of total baptism counts by year

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



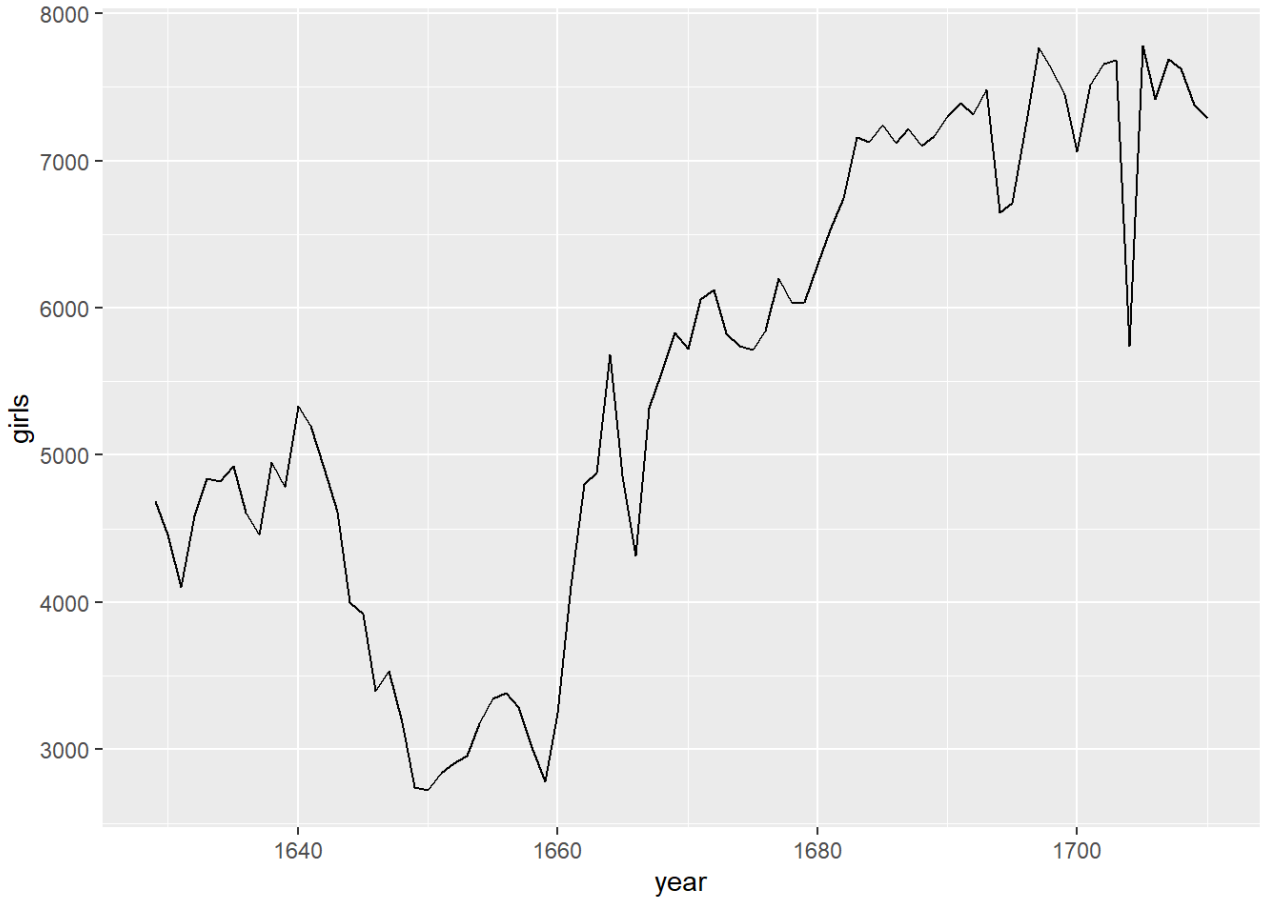
Hide

```
#create scatterplot of baptism counts (boys) by year

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



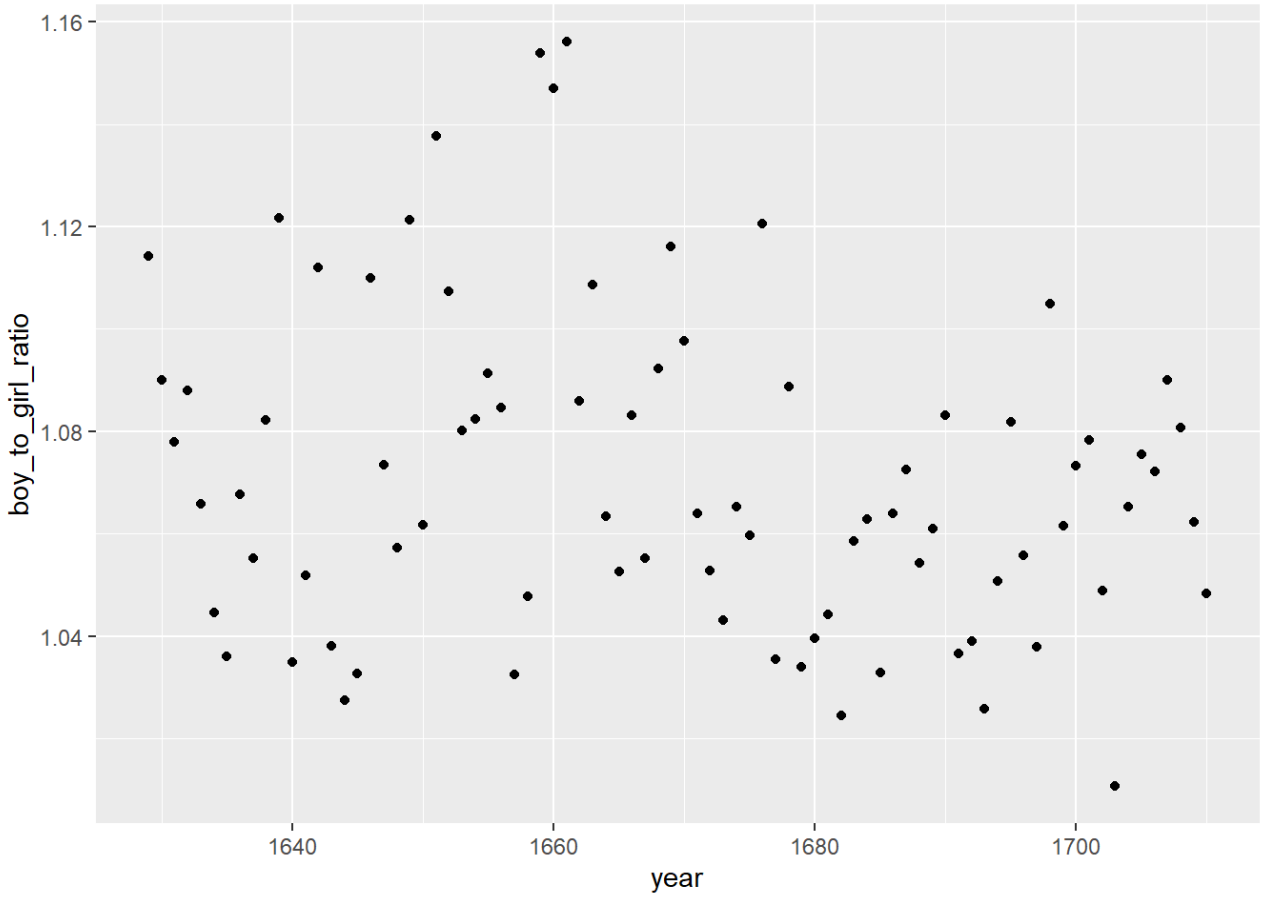
Hide

```
#create scatterplot of baptism counts (girls) count by year

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



Hide

```
# create new tbl column with ratio boys:girls

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

#create scatterplot for the ratio of boy:girl baptism counts by year

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



Hide

```
# create new tbl column for the ratio of boy baptisms (to total) each year

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

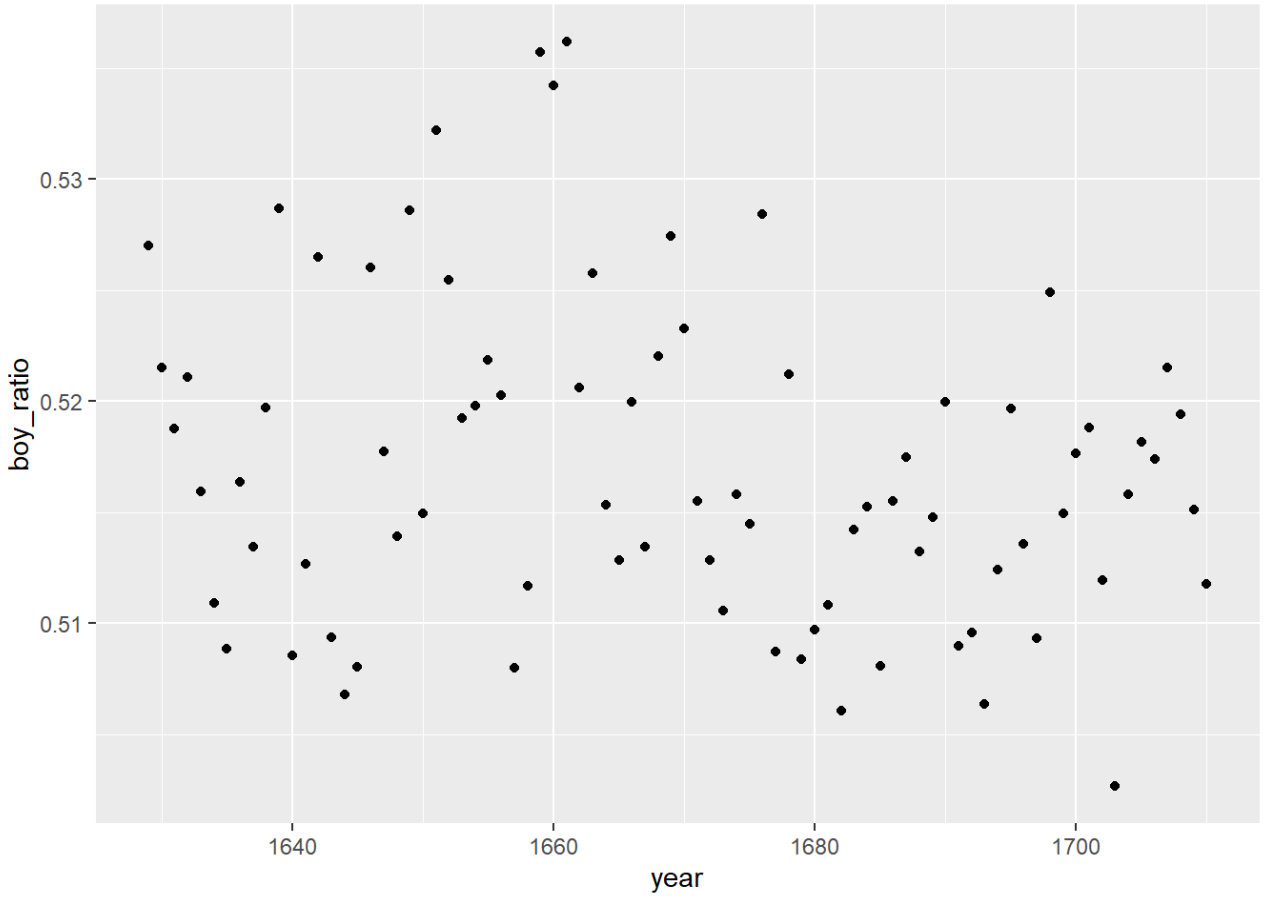
Exercise 3

It is difficult to discern a obvious trend(s) in the proportion of boys born over time. This owes, in part, to high inter-annual variation.

Note: I don't believe we have established that baptism counts accurately track male/female births year-to-year.

Hide

```
#create scatterplot for the ratio of boy baptism counts (to total) each year

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



Hide

```
# evaluate number of births of boys relative to girls each year (see also: Exercise 2)

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

# find the min and max count for boys

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

```
## # A tibble: 1 x 2
##     min   max
##   <int> <int>
## 1  2890  8426
```

Exercise 4

Included Years: 1940-202

DF dimensions: 63 rows, 3 cols

Column names: year, boys, girls

Hide

```
#preview data set

present
```

```
## # A tibble: 63 x 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

```
# return years column

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
```

Hide

```
# return dataframe dimensions

dim(present)
```

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

Hide

```
# return column names

names(present)
```

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

Exercise 5

Annual birth counts in the 'Present' data set are several orders of magnitude greater than annual baptism counts included in the Arbuthnot data set.

Each data set covers a similar period of time, ~60 yrs.

In the 'Present' data set, total birth counts increased by 1661327 between 1940 & 2002. Interim trends in birth counts were equivalent for boys and girls.

There was notable, but temporary, decline in birth rates that culminated in the mid-70s. I speculate that this may related to economic stresses related to the oil crisis, etc.

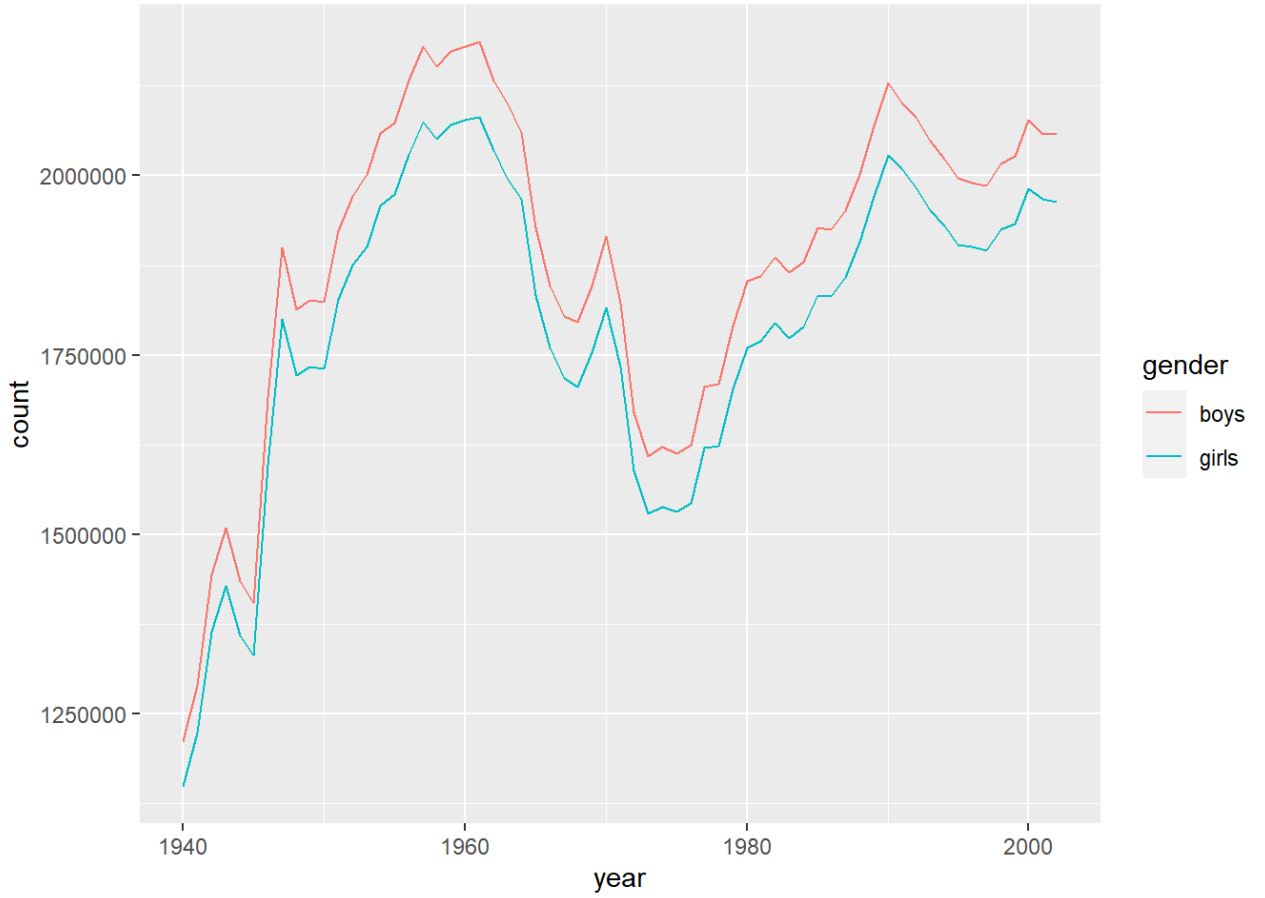
Hide

```
# return max and min birth counts for boys and girls over period of record

present%>%pivot_longer(cols=c(boys, girls), names_to = 'gender', values_to ='value')%
         >%group_by(gender)%>%summarize(min=min(value), max=max(value))
```

```
## # A tibble: 2 x 3
##   gender     min     max
## * <chr>    <dbl>   <dbl>
## 1 boys   1211684 2186274
## 2 girls  1148715 2082052
```

Hide

```
# plot birth count of boys vs girls by year

plot_gender <- present%>%pivot_longer(cols=c(boys, girls), names_to = 'gender', values
         _to ='count')

ggplot(plot_gender, aes(x=year, y=count, color=gender)) +geom_line()
```



Hide

```
# create new column for total birth counts, plot counts by year, return difference in
          total count between 1940 & 2002

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

plot_total <- ggplot(data=present, aes(x=year, y=total))+geom_line()

present$total%>%last()-present$total%>%first()
```

```
## [1] 1661327
```

Exercise 6

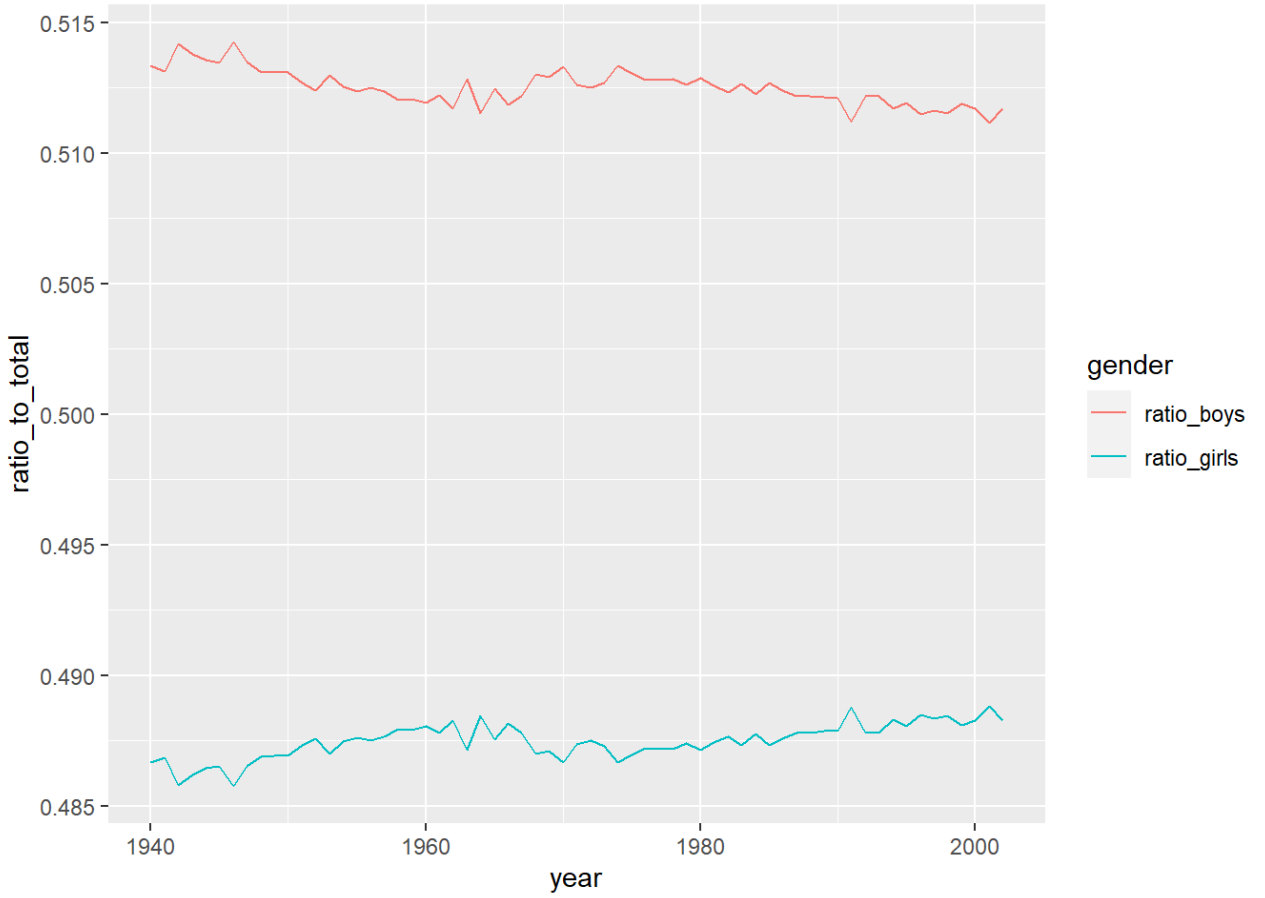
Arbuthnot's observation that boys are born in greater proportion than girls in the U.S. does hold true in the 'Present' dataset.

The offset in boy and girl birth counts (relative to total) is similar in magnitude across data sets.

However in the 'Present' data set, the proportion of male births (relative to total) decreases over time, while the opposite is true for females.

Hide

```
# create new column with ratio of boys:total births by year, repeat  for girls, plot r
        atios for each gender.

present<- present%>%mutate(ratio_boys=boys/total)%>%mutate(ratio_girls=girls/total)

plot_ratio <- present%>%pivot_longer(cols=c(ratio_boys, ratio_girls), names_to = 'gend
        er', values_to ='ratio_to_total')

ggplot(plot_ratio, aes(x=year, y=ratio_to_total, color=gender))+geom_line()
```



Exercise 7

The greatest total number of births in the U.S. (for the period of record) occurred in 1961.

Hide

```
present %>%arrange(desc(total))
```

```
## # A tibble: 63 x 6
##     year    boys   girls   total ratio_boys ratio_girls
##    <dbl>   <dbl>   <dbl>   <dbl>      <dbl>       <dbl>
##  1  1961 2186274 2082052 4268326      0.512       0.488
##  2  1960 2179708 2078142 4257850      0.512       0.488
##  3  1957 2179960 2074824 4254784      0.512       0.488
##  4  1959 2173638 2071158 4244796      0.512       0.488
##  5  1958 2152546 2051266 4203812      0.512       0.488
##  6  1962 2132466 2034896 4167362      0.512       0.488
##  7  1956 2133588 2029502 4163090      0.513       0.487
##  8  1990 2129495 2028717 4158212      0.512       0.488
##  9  1991 2101518 2009389 4110907      0.511       0.489
## 10  1963 2101632 1996388 4098020      0.513       0.487
## # ... with 53 more rows
```