AndrewBowen_Data606_LAb1

Andrew Bowen

2022-08-26

```
## Loading required package: airports

## Loading required package: cherryblossom

## Loading required package: usdata

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

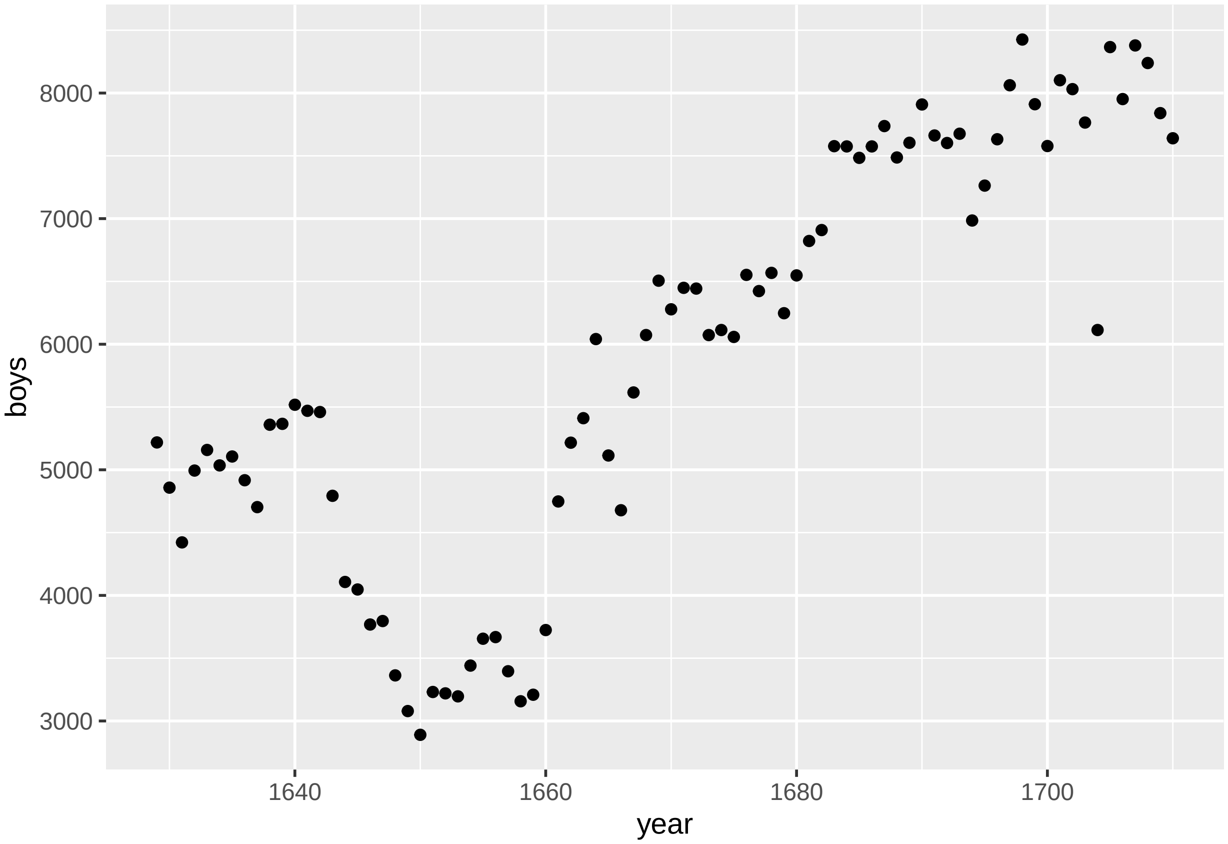
```
data('arbuthnot', package='openintro')

head(arbuthnot, 10)
```

```
## # A tibble: 10 x 3
##     year  boys girls
##    <int> <int> <int>
##  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
```

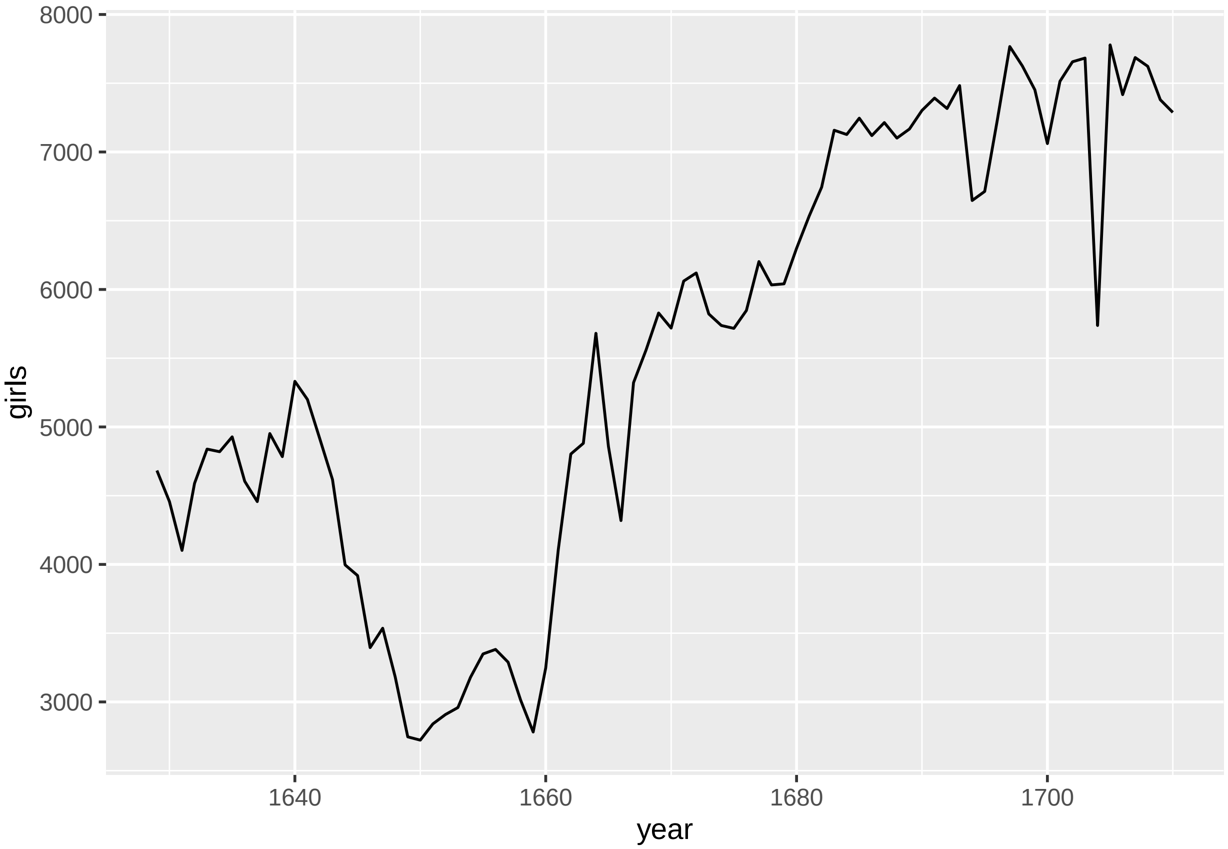
Plotting our DF

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



Same plot as above but in line format

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



```
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 `total` field to dataframe

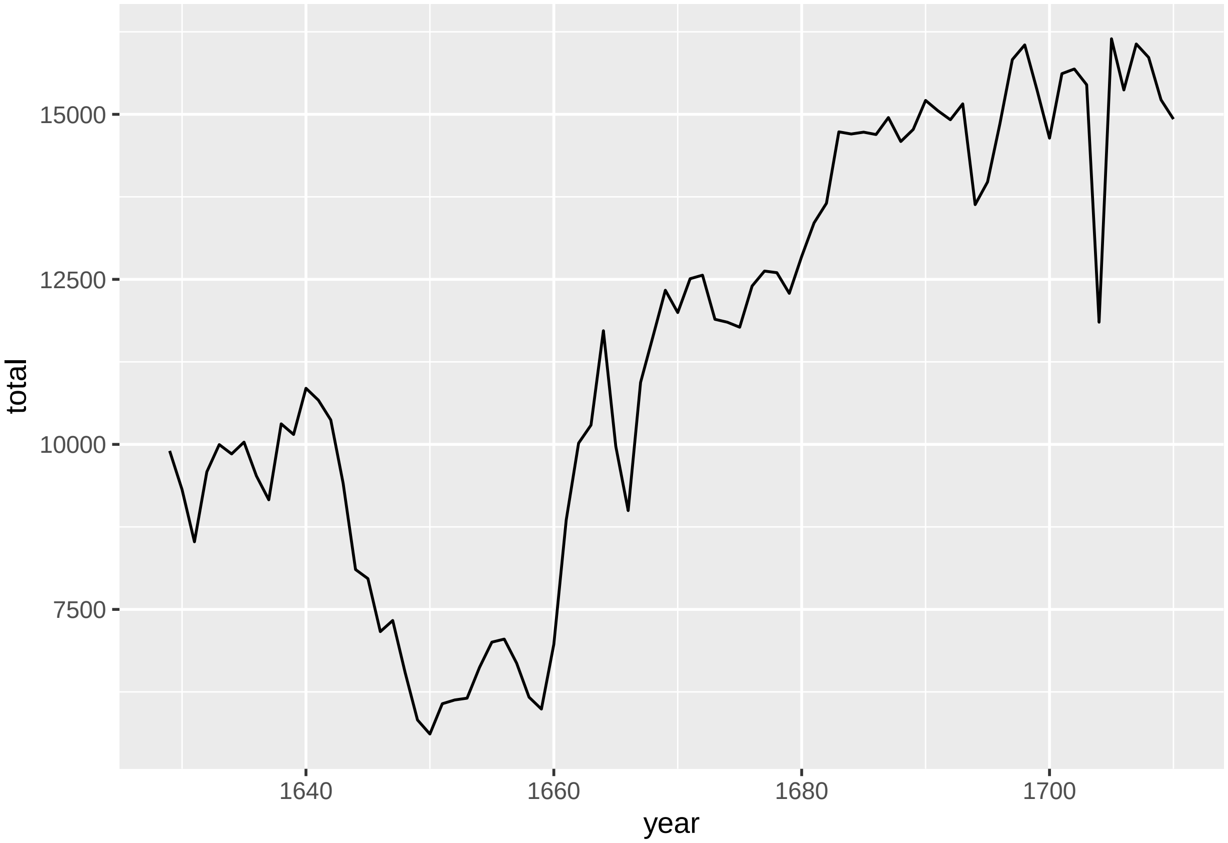
```
arbuthnot <- arbuthnot %>% mutate(total = boys + girls)

head(arbuthnot, 10)
```

```
## # A tibble: 10 x 4
##     year  boys girls total
##    <int> <int> <int> <int>
##  1  1629  5218  4683  9901
##  2  1630  4858  4457  9315
##  3  1631  4422  4102  8524
##  4  1632  4994  4590  9584
##  5  1633  5158  4839  9997
##  6  1634  5035  4820  9855
##  7  1635  5106  4928 10034
##  8  1636  4917  4605  9522
##  9  1637  4703  4457  9160
## 10  1638  5359  4952 10311
```

Plotting `total` over time

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



Calculationg boy-to-girl ratio

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

Calculating the ratio of boys baptized to total.

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

Finding years where more boys were present than girls. Adding in a boolean flag column to represent

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

Let's take a look at our DF with all our added columns!

```
head(arbuthnot, 10)
```

```
## # A tibble: 10 x 7
##     year  boys girls total boy_to_girl_ratio boy_ratio more_boys
##    <int> <int> <int> <int>             <dbl>     <dbl> <lgl>
##  1  1629  5218  4683  9901              1.11     0.527 TRUE
##  2  1630  4858  4457  9315              1.09     0.522 TRUE
##  3  1631  4422  4102  8524              1.08     0.519 TRUE
##  4  1632  4994  4590  9584              1.09     0.521 TRUE
##  5  1633  5158  4839  9997              1.07     0.516 TRUE
##  6  1634  5035  4820  9855              1.04     0.511 TRUE
##  7  1635  5106  4928 10034              1.04     0.509 TRUE
##  8  1636  4917  4605  9522              1.07     0.516 TRUE
##  9  1637  4703  4457  9160              1.06     0.513 TRUE
## 10  1638  5359  4952 10311              1.08     0.520 TRUE
```

More Practice

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

```
head(present, 10)
```

```
## # A tibble: 10 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
```

Question 1

Let’s see our data range in the year column first (using `min` & `max`)

```
print(min(present$year))
```

```
## [1] 1940
```

```
print(max(present$year))
```

```
## [1] 2002
```

Finding out our data frame dimensions (nrows x ncols)

```
print(ncol(present))
```

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

```
print(nrow(present))
```

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

Getting out dataframe columns with the built-in `colnames` function

```
colnames(present)
```

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

Question 2 Going to use the median count for boys and girls from each data set (`present` vs `arbuthnot`) to compare magnitudes of counts in each data set

```
p_boys_med = median(present$boys)
p_girls_med = median(present$girls)
```

```r
a_boys_med = median(arbuthnot$boys)
a_girls_med = median(arbuthnot$girls)

# Calculating ratio of present boy/girl counts
boys_ratio = p_boys_med / a_boys_med
girls_ratio = p_girls_med / a_girls_med
print(boys_ratio)
```

```
## [1] 316.955
```

```r
print(girls_ratio)
```

```
## [1] 320.3356
```

It looks like the median present day birth counts are ~320 times higher than the birth counts listed in our `arbuthnot` dataset. We used median counts to summarize the data set, so it won't be exactly this ratio for the whole data set, but modern birth counts are significantly higher.

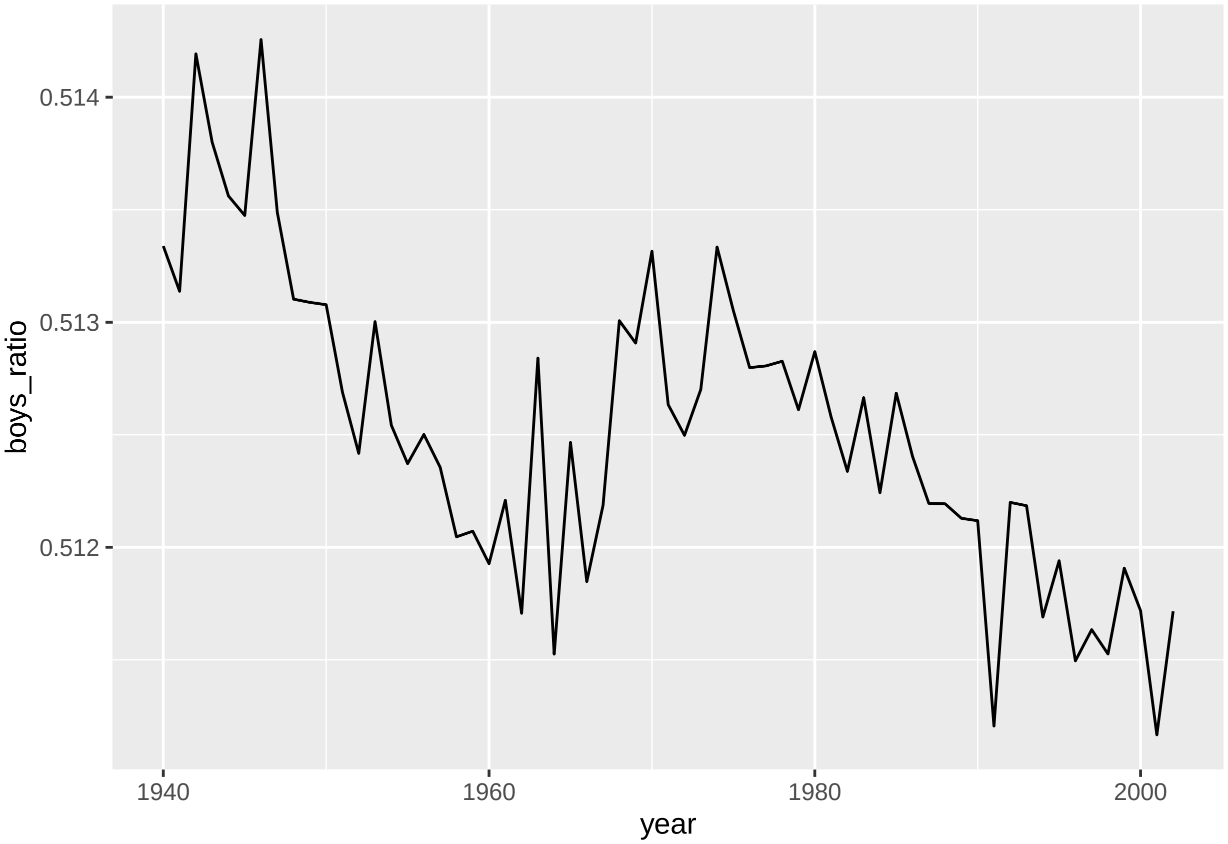
Question 3 Setting up our boy-girl ratio column

```r
present <- present %>% mutate(total = boys + girls)
present <- present %>% mutate(boys_ratio = boys / total)
head(present, 10)
```

```
## # A tibble: 10 x 5
##     year    boys   girls   total boys_ratio
##    <dbl>   <dbl>   <dbl>   <dbl>      <dbl>
##  1  1940 1211684 1148715 2360399      0.513
##  2  1941 1289734 1223693 2513427      0.513
##  3  1942 1444365 1364631 2808996      0.514
##  4  1943 1508959 1427901 2936860      0.514
##  5  1944 1435301 1359499 2794800      0.514
##  6  1945 1404587 1330869 2735456      0.513
##  7  1946 1691220 1597452 3288672      0.514
##  8  1947 1899876 1800064 3699940      0.513
##  9  1948 1813852 1721216 3535068      0.513
## 10  1949 1826352 1733177 3559529      0.513
```

Let's plot the ratio of boys born over time in our `present` dataset:

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



While the ratio of boys born has stayed over 50%, it is experiencing a downward trend over time since 1940, the beginning of our dataset. The observation of boys being born more than girls from the `arbuthnot` dataset does hold up, but has decreased since 1940 in the US.

Question 4

```
# Truncating output to 10 rows for readability
head(present %>% arrange(desc(total)), 10)
```

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

We see the highest number of total births in the US come in 1961 with 4268326 total births (boys & girls). It's interesting to note that 8 of the top 10 years in terms of total births came during the baby boom years in the post-war era.

Checking out the help function (`?`) built into R objects.

```
?present
```