Lab1

Alice Ding

2023-02-01

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.4.0      v purrr   1.0.1
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.3.0      v stringr 1.5.0
## v readr   2.1.3      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## Loading required package: airports
##
## Loading required package: cherryblossom
##
## Loading required package: usdata
```

Exercise 1

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

```
arbuthnot$girls
```

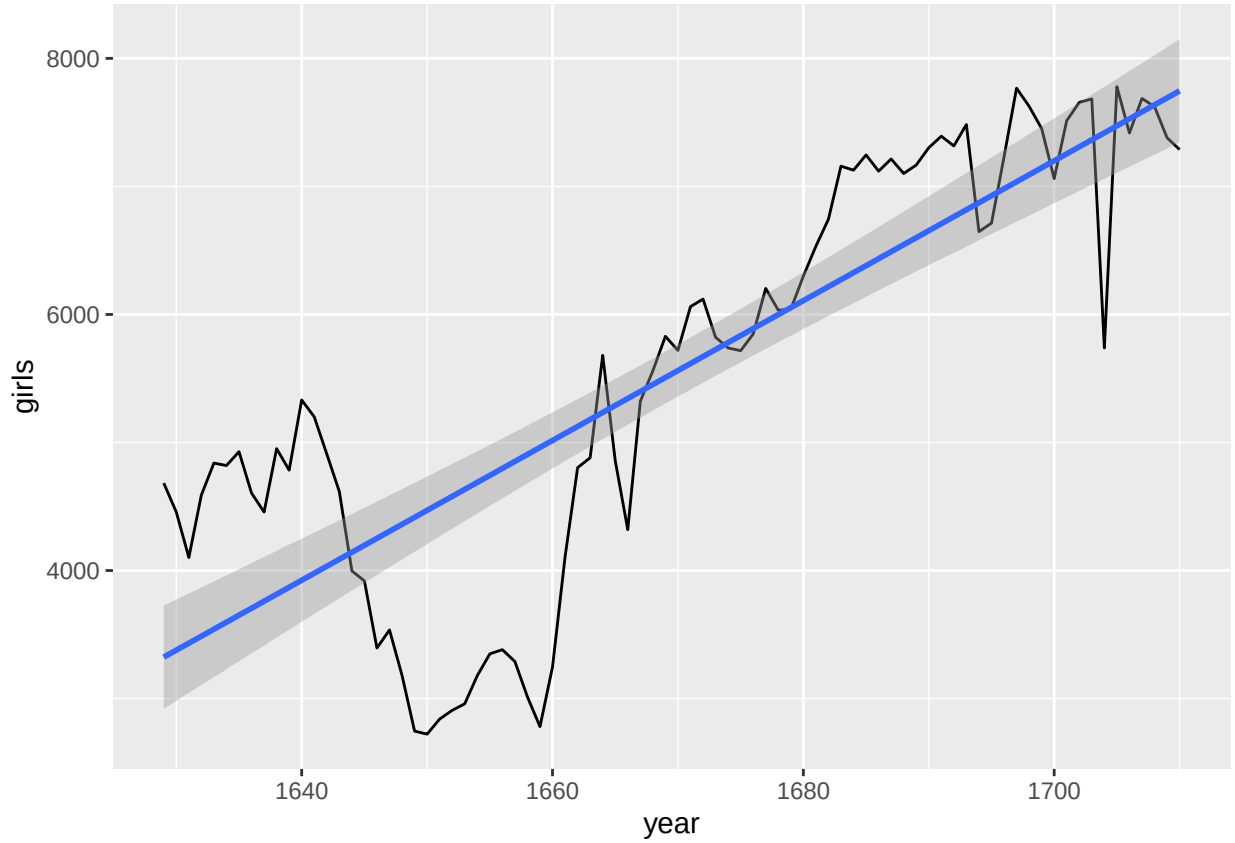
```
##  [1] 4683 4457 4102 4590 4839 4820 4928 4605 4457 4952 4784 5332 5200 4910 4617
## [16] 3997 3919 3395 3536 3181 2746 2722 2840 2908 2959 3179 3349 3382 3289 3013
## [31] 2781 3247 4107 4803 4881 5681 4858 4319 5322 5560 5829 5719 6061 6120 5822
## [46] 5738 5717 5847 6203 6033 6041 6299 6533 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
```

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.)

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

```
## `geom_smooth()` using formula = 'y ~ x'
```

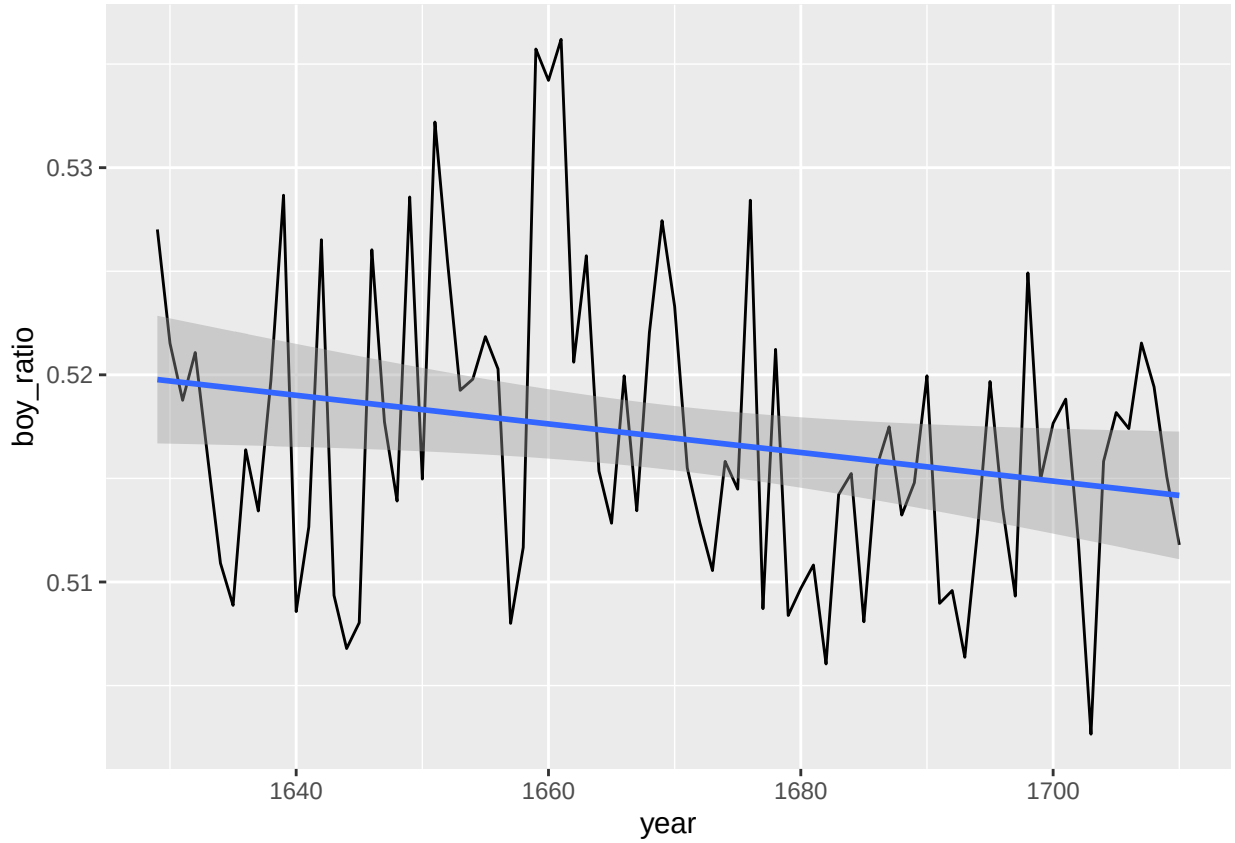


It seems like the number of girls baptised was somewhat consistent starting in 1629 before crashing starting in 1640. This would suggest maybe there was a slower birth rate, and this lasted until around 1660 before ticking upwards in a consistent upwards trend until hitting a sort of plateau in the mid 1690s. There were some outlier years (the mid 1660s had a pretty huge crash), but otherwise baptisms were in the 7000s starting in the early 1680s, minus one huge outlier in the early 1700s.

Overall, it seems like the number of girls getting baptised has almost doubled from 1629-1700s. If we take the lowest values and compare them to the highest, then they did double. ## Exercise 3 ### Now, generate a plot of the proportion of boys born over time. What do you see?

```
arbuthnot <- arbuthnot %>%
  mutate(total = boys + girls)
arbuthnot <- arbuthnot %>%
  mutate(boy_ratio = boys / total)
ggplot(data = arbuthnot, aes(x = year, y = boy_ratio)) +
  geom_line() +
  geom_smooth(method=lm)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



I've added a trend line here to show that the proportion of boys being born over time is slowly declining. We see it had an average of ~0.52 to start, but it's been slowly moving towards ~0.515 as time has passed – it's still above 50% so more boys are being born than girls, but it's almost decreased by 0.01 over the ~80 years for this data set.

Exercise 4

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

```
dim(present)
```

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

```
glimpse(present)
```

```
## Rows: 63
## Columns: 3
## $ year  <dbl> 1940, 1941, 1942, 1943, 1944, 1945, 1946, 1947, 1948, 1949, 1950~
## $ boys  <dbl> 1211684, 1289734, 1444365, 1508959, 1435301, 1404587, 1691220, 1~
## $ girls <dbl> 1148715, 1223693, 1364631, 1427901, 1359499, 1330869, 1597452, 1~
```

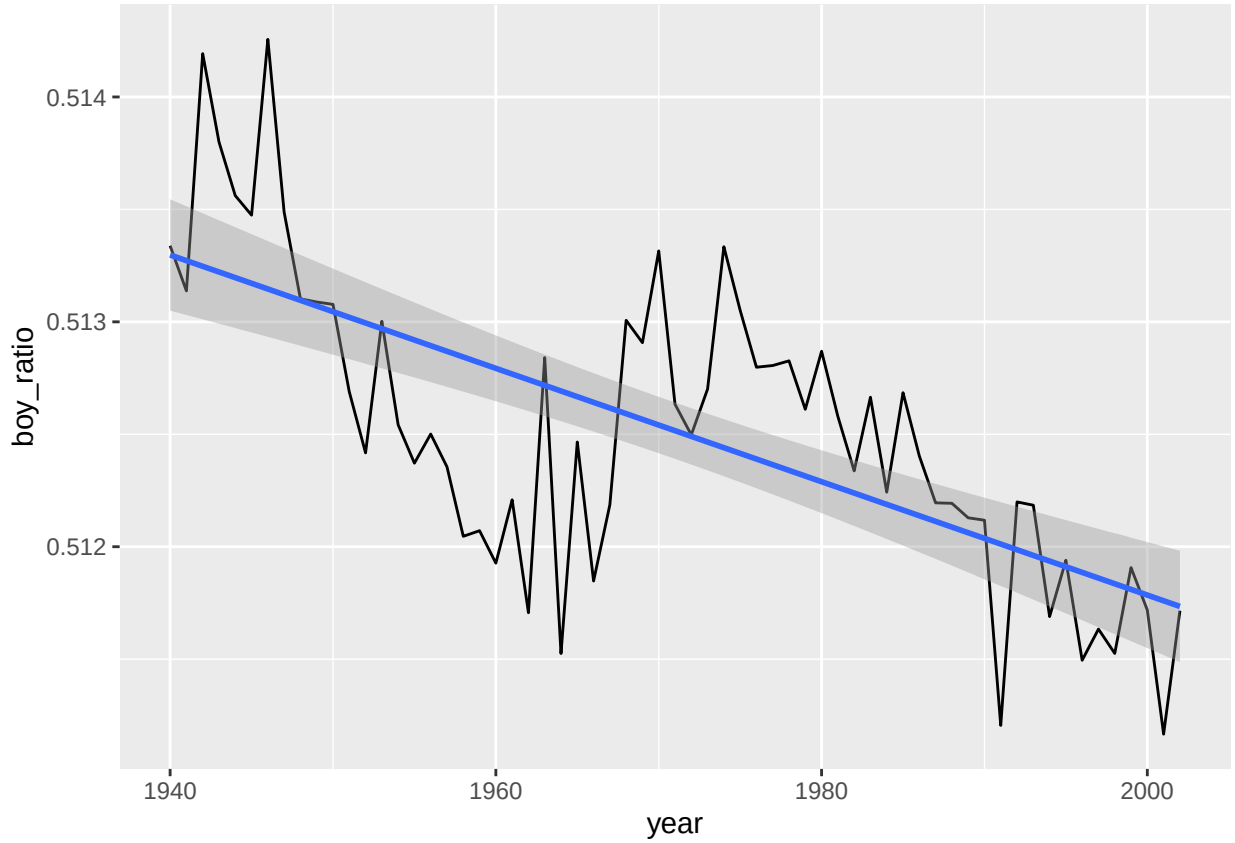
```
present %>%
  summarize(min = min(year), max = max(year))
```

```
## # A tibble: 1 x 2
##     min   max
##   <dbl> <dbl>
## 1  1940  2002
```

This data set includes years from 1940 to 2002. The dataframe has 3 columns and the variable names are `year`, `boys`, and `girls`. ## Exercise 5 ### How do these counts compare to Arbuthnot's? Are they of a similar magnitude? These counts are much larger compared to Arbuthnot's – just as a basic observation, Arbuthnot's had their boys and girls counts have 4 digits ranging from 3-7k, Present's numbers are in the millions for both genders. ## 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.

```
present <- present %>%
  mutate(total = boys + girls)
present <- present %>%
  mutate(boy_ratio = boys / total)
ggplot(data = present, aes(x = year, y = boy_ratio)) +
  geom_line() +
  geom_smooth(method=lm)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



Yes, it does look like boys are being born in greater proportion for the US as well, albeit at a slightly smaller difference. We also see this proportion declining, meaning girls are likely being born at an increasing rate, although this is still above 50% so boys are more popular nonetheless. ## Exercise 7 ### In what year did we see the most total number of births in the U.S.?

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

```
## # A tibble: 63 x 5
##     year    boys   girls   total boy_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
## # ... with 53 more rows
```

1961 had 4,268,326 births.