Lab 1: Intro to R

Leticia Salazar

2021-09-04

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

Exercise 1

Counts of girls baptized:

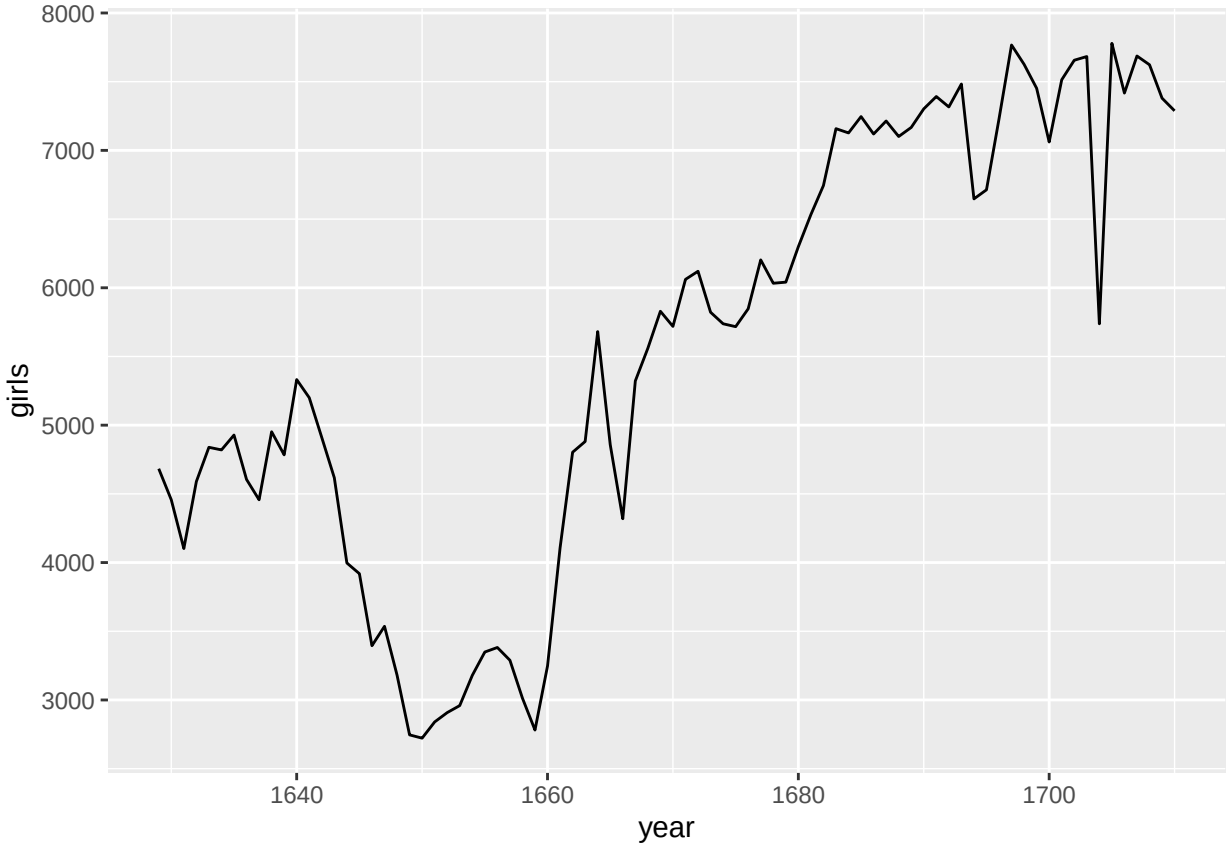
```
arbuthnot$girls
```

```
##  [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
```

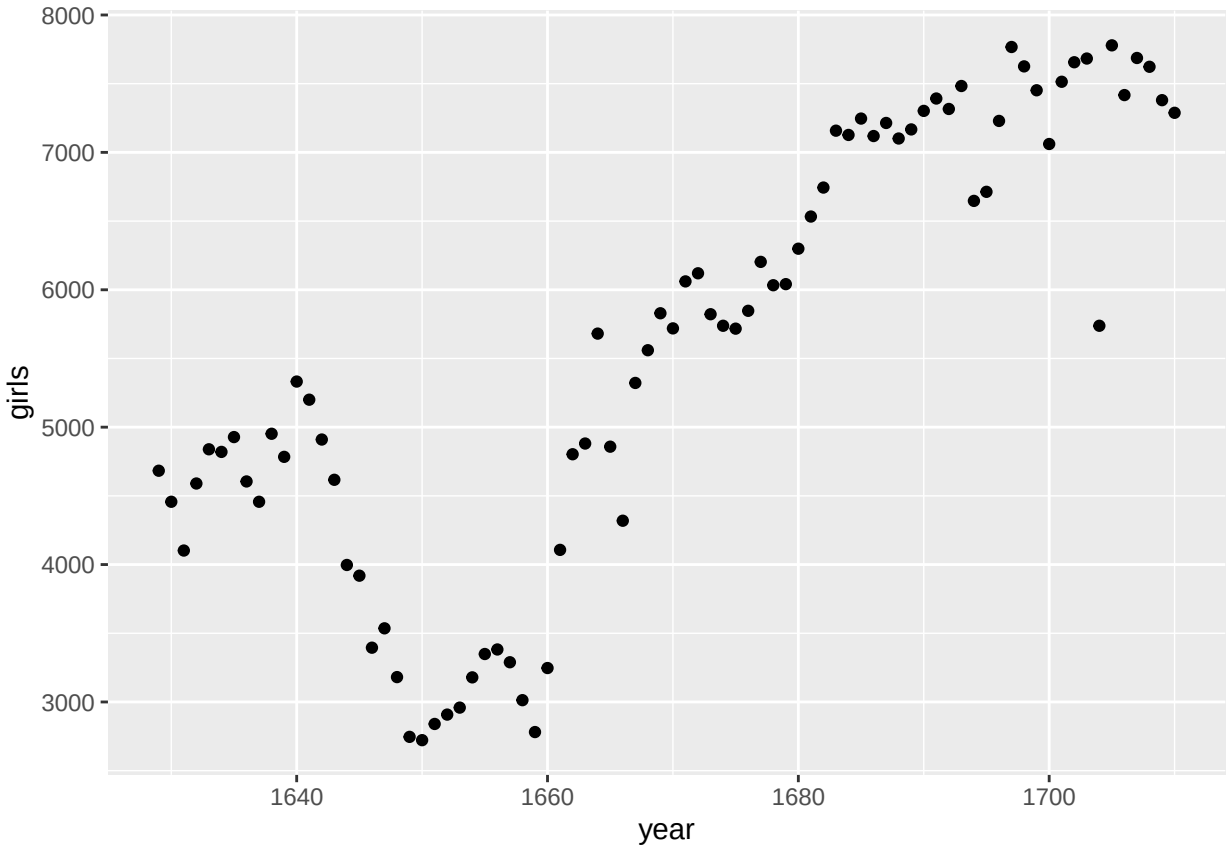
Exercise 2

There is a peak around 1660 with the number of girls being baptized increases with a slight decrease after the 1700s. Overall, it's an upward trend in girls being baptized.

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




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

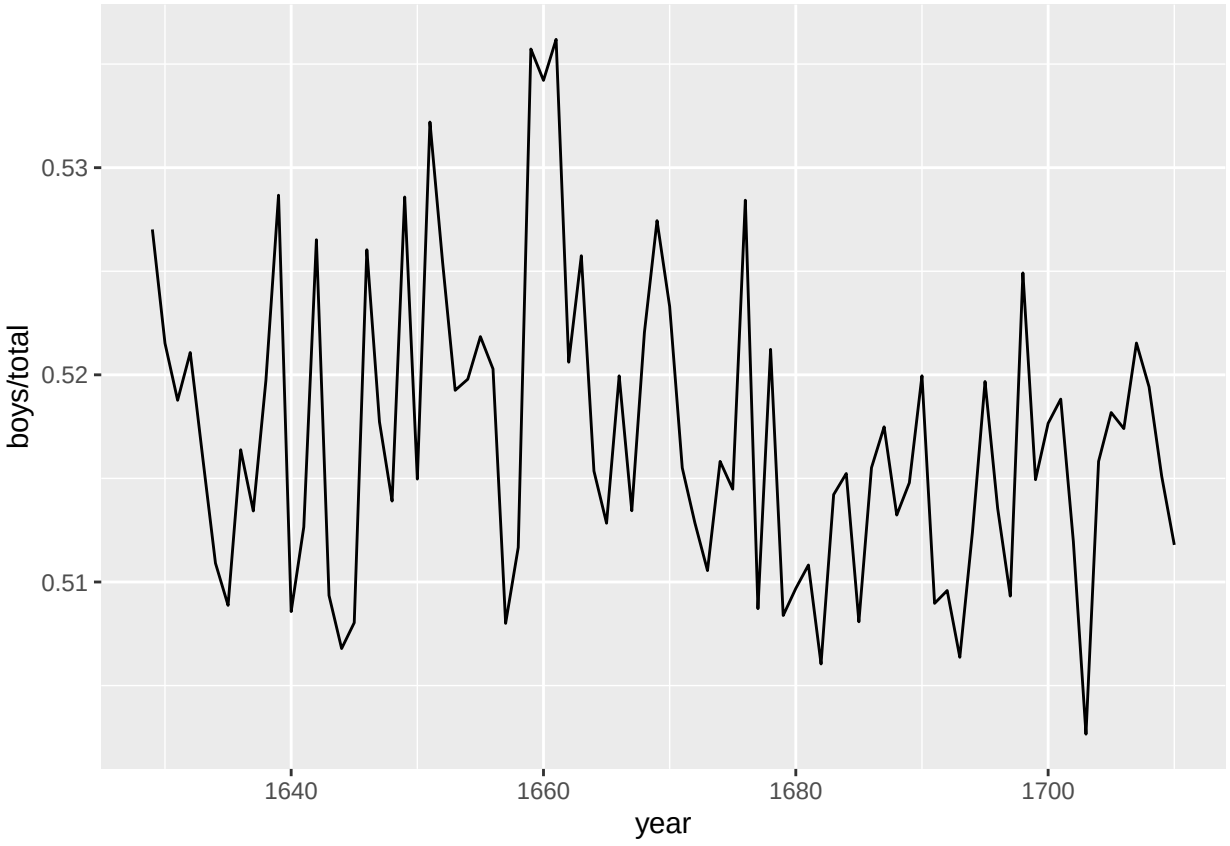


Exercise 3

The boys born over time are slightly above 50% up until early 1700s where there's a sudden drop. Overall, the birth of boys has significantly increased.

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

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



Exercise 4

The years included in this data set are from 1940 to 2002. The dimensions of the data set are 63, 3 and the column names are "year", "boys", "girls".

```
data(present)

dim(present)
```

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

```
range(present$year)
```

```
## [1] 1940 2002
```

```
names(present)
```

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

Exercise 5

Comparing the arbuthnot's and present data set they are completely different. The median for arbuthnot's data set is 11813 where as the one for present is 3756547.

```
summary(arbuthnot)
```

```
##       year           boys          girls          total
##  Min.   :1629   Min.   :2890   Min.   :2722   Min.   : 5612
##  1st Qu.:1649   1st Qu.:4759   1st Qu.:4457   1st Qu.: 9199
##  Median :1670   Median :6073   Median :5718   Median :11813
##  Mean   :1670   Mean   :5907   Mean   :5535   Mean   :11442
##  3rd Qu.:1690   3rd Qu.:7576   3rd Qu.:7150   3rd Qu.:14723
##  Max.   :1710   Max.   :8426   Max.   :7779   Max.   :16145
```

```
summary(present)
```

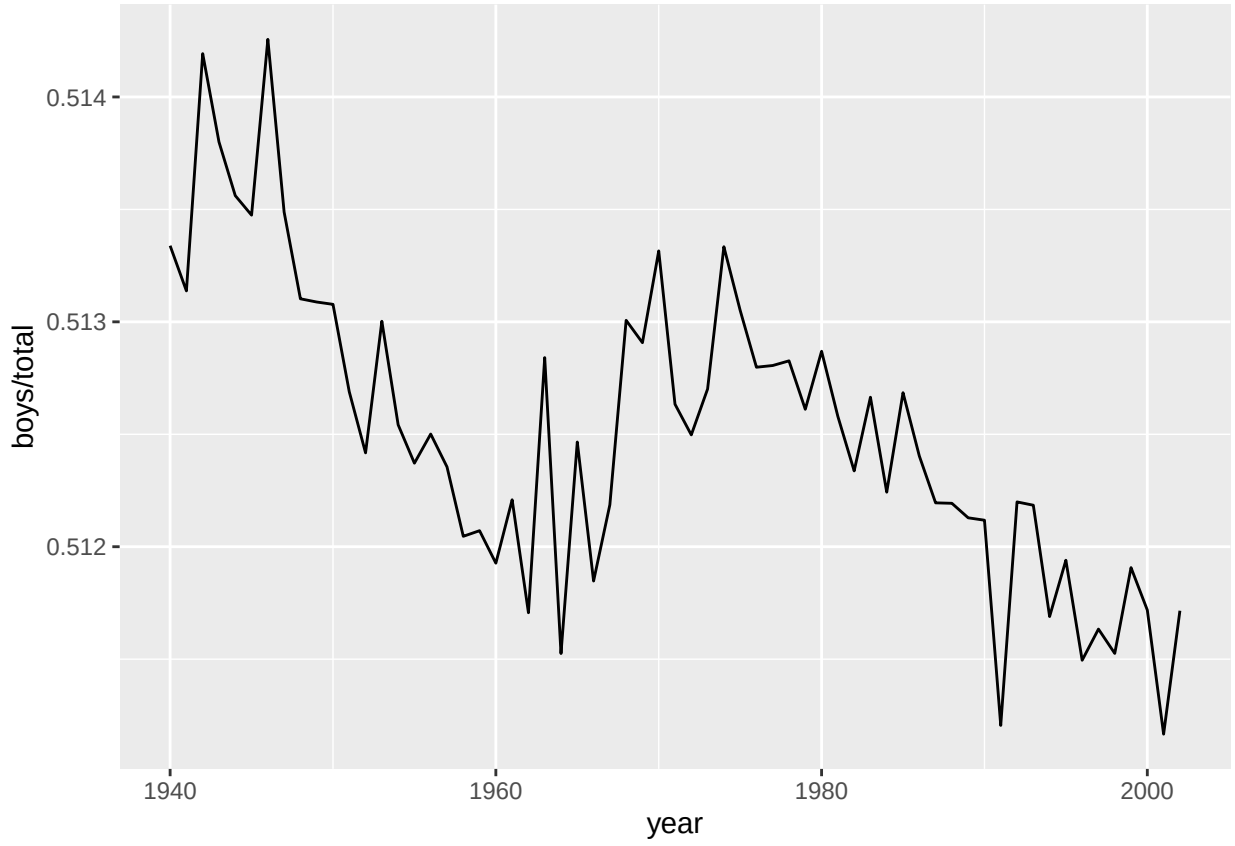
```
##       year           boys             girls
##  Min.   :1940   Min.   :1211684   Min.   :1148715
##  1st Qu.:1956   1st Qu.:1799857   1st Qu.:1711404
##  Median :1971   Median :1924868   Median :1831679
##  Mean   :1971   Mean   :1885600   Mean   :1793915
##  3rd Qu.:1986   3rd Qu.:2058524   3rd Qu.:1965538
##  Max.   :2002   Max.   :2186274   Max.   :2082052
```

Exercise 6

Looking at the graph, there is a decrease in boys being born in the US compared to Arbuthnot's observation. You can see that between 1940 and 1945 there's still a peak but after 1950's there's a decline. From 1965 to 1980 there's a slight increase with a huge drop afterwards.

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

ggplot(data = present, aes(x = year, y = boys/total)) +
  geom_line()
```



Exercise 7

Based on this data, the most number of births in the US was in 1961.

```
present %>%
  mutate(total = boys + girls) %>%
  arrange(desc(total))
```

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