Lab 1: Intro to R

Avery Davidowitz

2022-09-04

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

Exercise 1

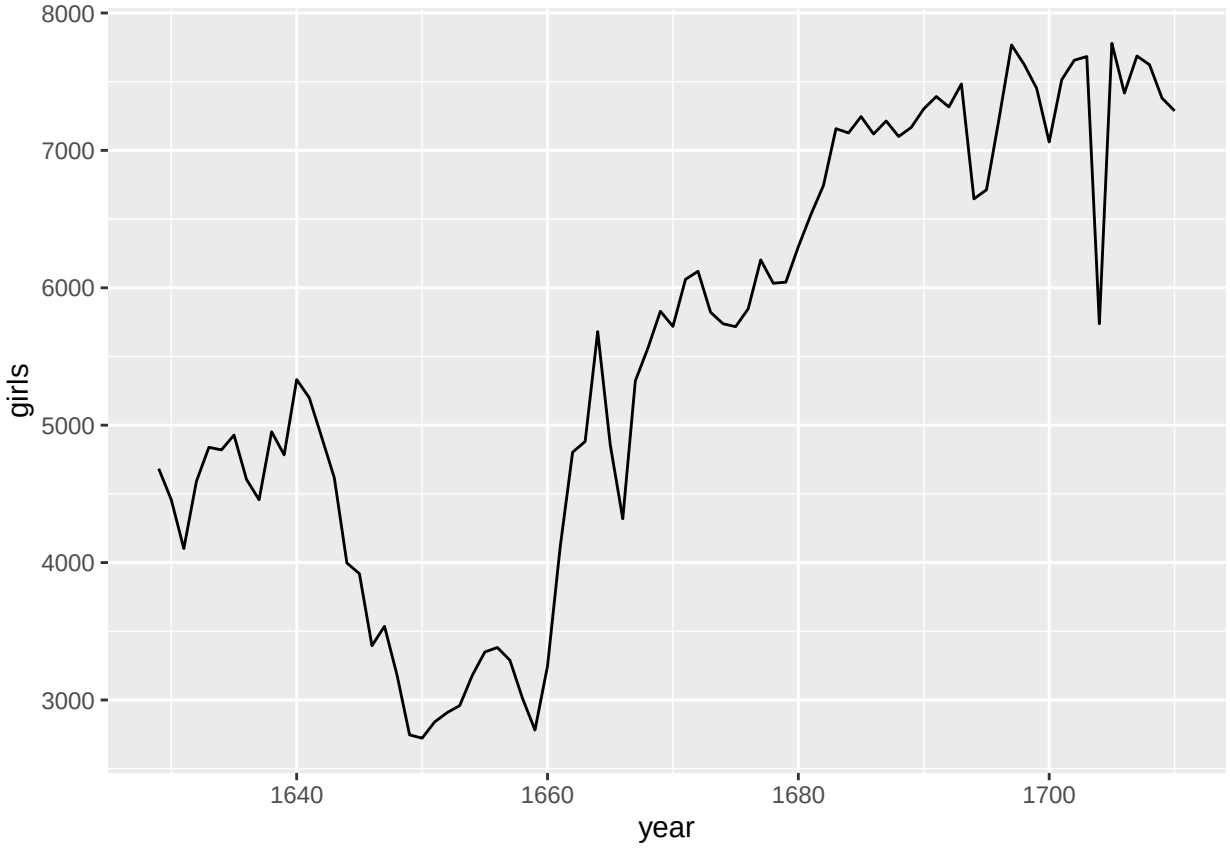
```
arbuthnot$girls
```

```
##  [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

Excluding a period of sharp decline centered around 1645 the number of girls baptized shows a steady slow increase as shown in the plot below.

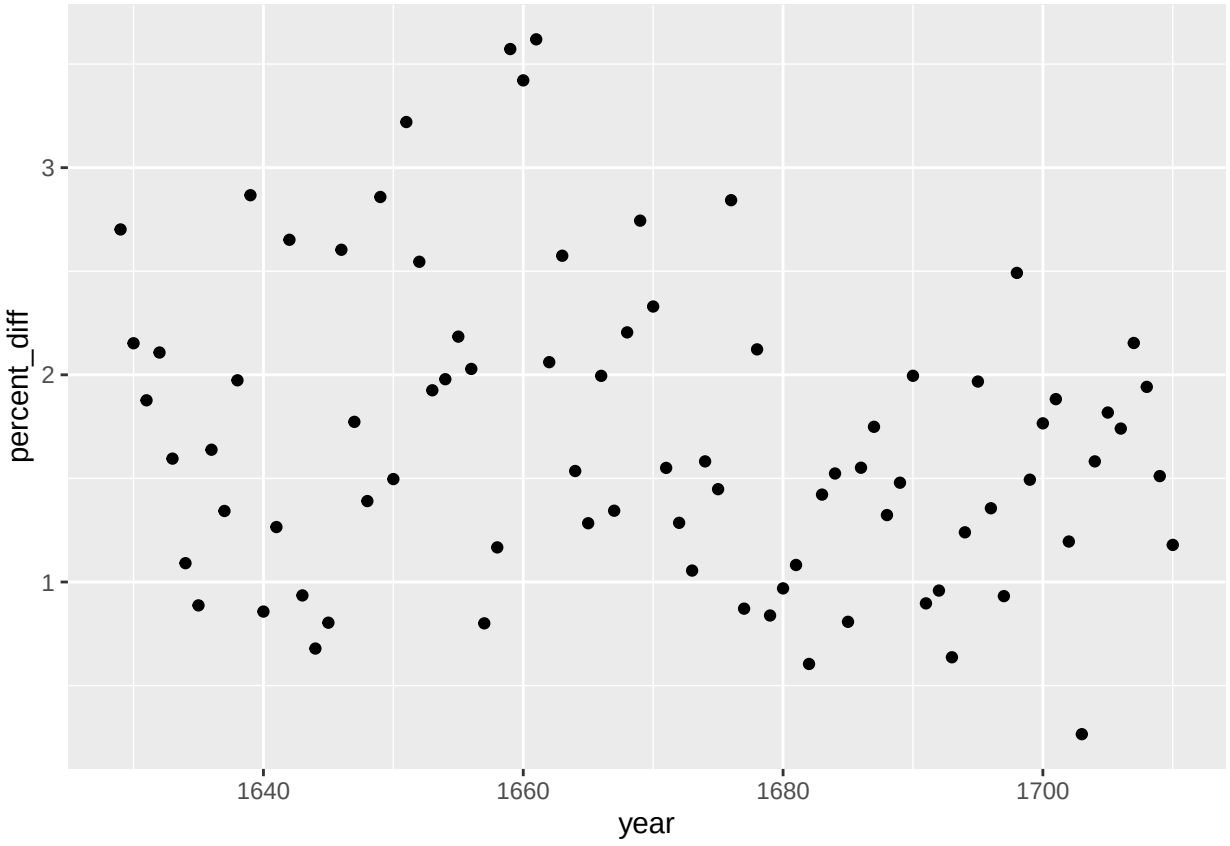
```
ggplot(data = arbuthnot, aes(x = year, y = girls)) +
  geom_line()
```



Exercise 3

The proportion of boys born over time is consistently higher than girls. Generally it vacillates between .5 and 3.5 percent higher than girls born as shown in the plot below.

```
arbuthnot <- arbuthnot %>%
  mutate(total = boys + girls)
arbuthnot <- arbuthnot %>%
  mutate(percent_diff = 100 * abs(boys - girls) / total / 2)
ggplot(data = arbuthnot, aes(x = year, y = percent_diff)) +
  geom_point()
```



Exercise 4

The data frame is 82 rows by 3 columns: year, boys and girls. The year spans from 1940 to 2002.

```
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
```

Exercise 5

As demonstrated by the average total number of children counted in a given year in each data frame, the US births were over 300 times higher than the London births.

```r
present <- present %>%
  mutate(total = boys + girls)
present %>%
  summarize(average = mean(total))
```

```
## # A tibble: 1 x 1
##    average
##      <dbl>
## 1 3679515.
```

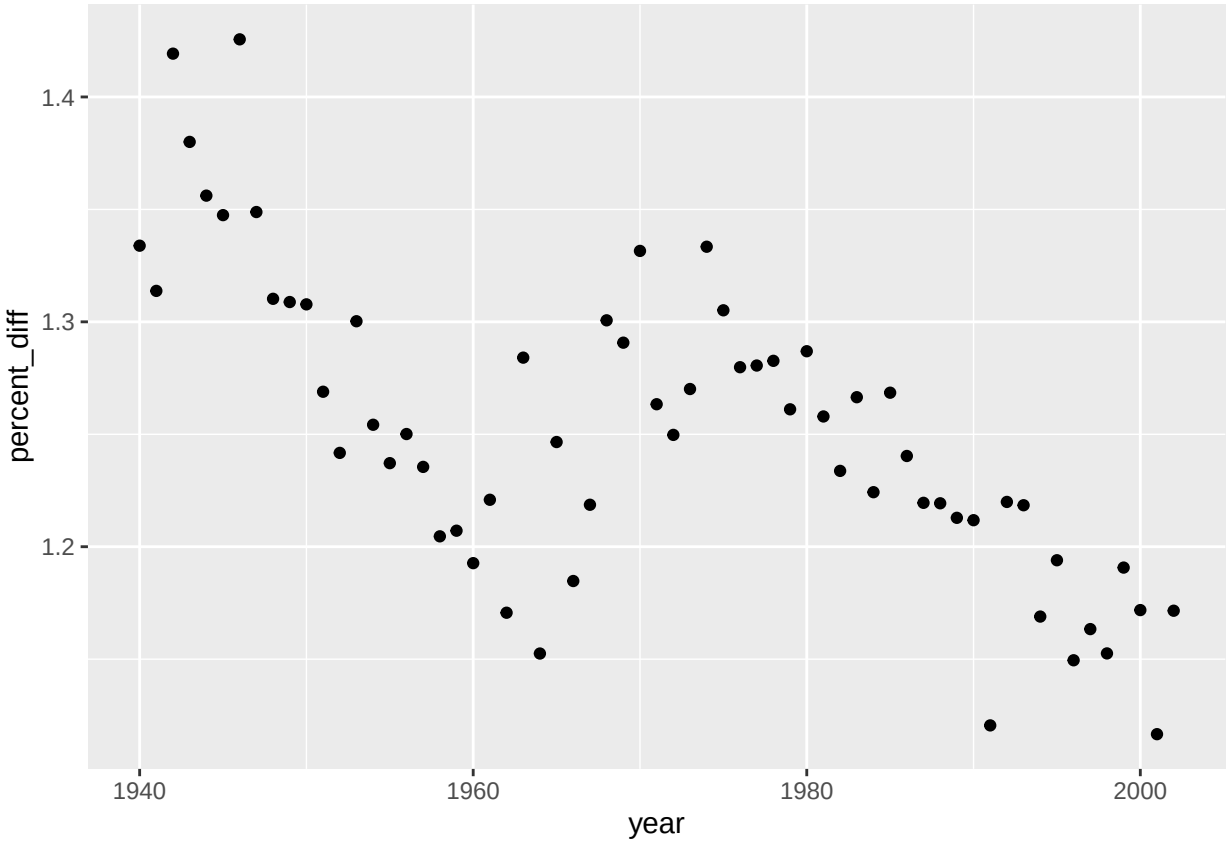
```r
arbuthnot %>%
  summarize(average = mean(total))
```

```
## # A tibble: 1 x 1
##   average
##     <dbl>
## 1  11442.
```

Exercise 6

Similar to the Arbuthnot data frame, boys are consistently born more commonly. However, in this data the difference in representation is noticeably less spread and stays between a 1% and 1.5% difference.

```r
present <- present %>%
  mutate(total = boys + girls)
present <- present %>%
  mutate(percent_diff = 100 * abs(boys - girls) / total / 2)
ggplot(data = present, aes(x = year, y = percent_diff)) +
  geom_point()
```



Exercise 7

1961 saw the most total births in the US between 1940 and 2002.

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

```
## # A tibble: 63 x 5
##     year    boys   girls   total percent_diff
##    <dbl>   <dbl>   <dbl>   <dbl>        <dbl>
##  1  1961 2186274 2082052 4268326         1.22
##  2  1960 2179708 2078142 4257850         1.19
##  3  1957 2179960 2074824 4254784         1.24
##  4  1959 2173638 2071158 4244796         1.21
##  5  1958 2152546 2051266 4203812         1.20
##  6  1962 2132466 2034896 4167362         1.17
##  7  1956 2133588 2029502 4163090         1.25
##  8  1990 2129495 2028717 4158212         1.21
##  9  1991 2101518 2009389 4110907         1.12
## 10  1963 2101632 1996388 4098020         1.28
## # ... with 53 more rows
```