Lab 1: Intro to R

Jiaxin Zheng

09/06/2024

```
library(tidyverse)
library(openintro)
library(ggplot2)
library(dplyr)
```

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
```

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

```
## # A tibble: 82 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
## # i 72 more rows
```

```
glimpse(arbuthnot)
```

```
## Rows: 82
## Columns: 3
```

```
## $ year  <int> 1629, 1630, 1631, 1632, 1633, 1634, 1635, 1636, 1637, 1638, 1639~
## $ boys  <int> 5218, 4858, 4422, 4994, 5158, 5035, 5106, 4917, 4703, 5359, 5366~
## $ girls <int> 4683, 4457, 4102, 4590, 4839, 4820, 4928, 4605, 4457, 4952, 4784~
```

```
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
```

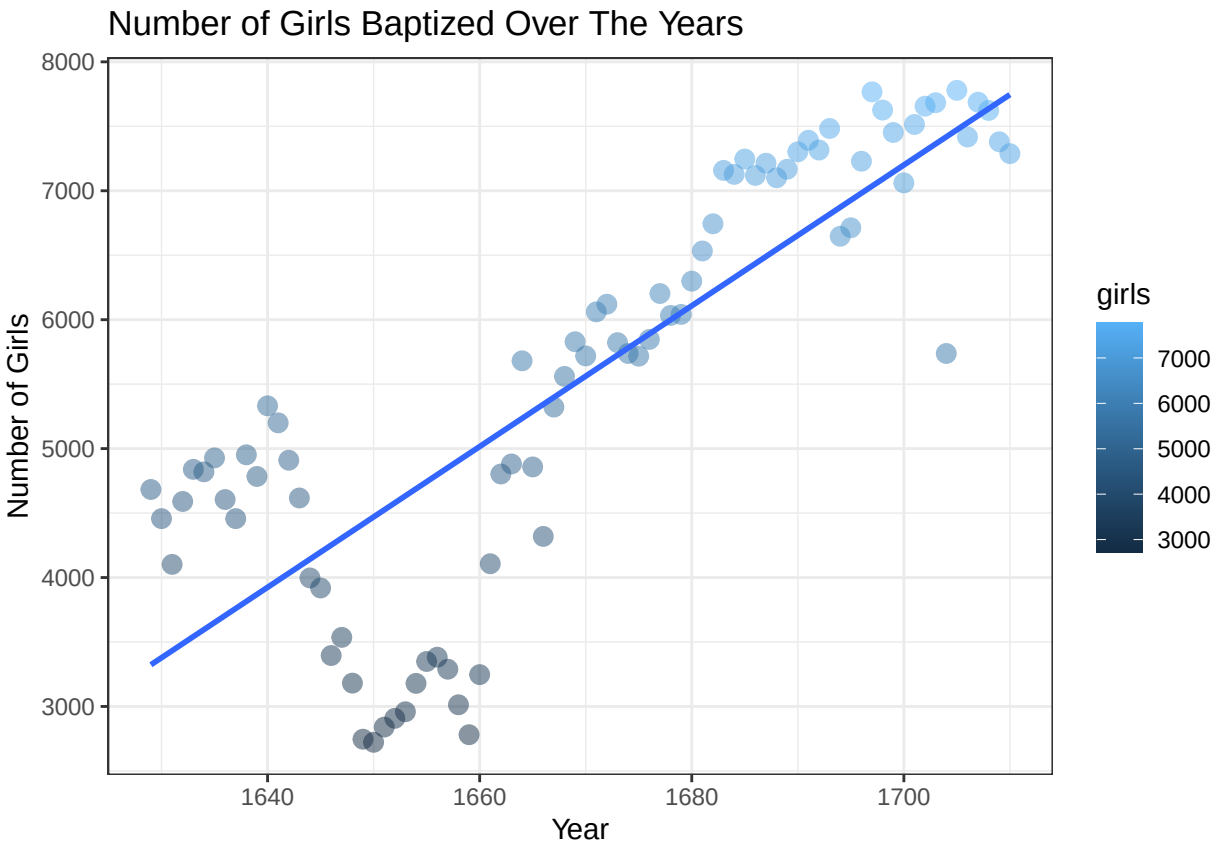
Exercise 2

There has decrease number in year 1640-1660. And huge increase number of girls baptized during 1680-1700.

```
# Insert code for Exercise 2 here
ggplot(data = arbuthnot, aes(x = year, y = girls,
                             colour= girls)) +
  geom_point(size=3, alpha=0.5)+
  geom_smooth(method = lm,
              se=F)+
  labs(title="Number of Girls Baptized Over The Years",
     x="Year",
     y="Number of Girls")+
  theme_bw()
```

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

```
## Warning: The following aesthetics were dropped during statistical transformation:
## colour.
## i This can happen when ggplot fails to infer the correct grouping structure in
##   the data.
## i Did you forget to specify a `group` aesthetic or to convert a numerical
##   variable into a factor?
```



Exercise 3

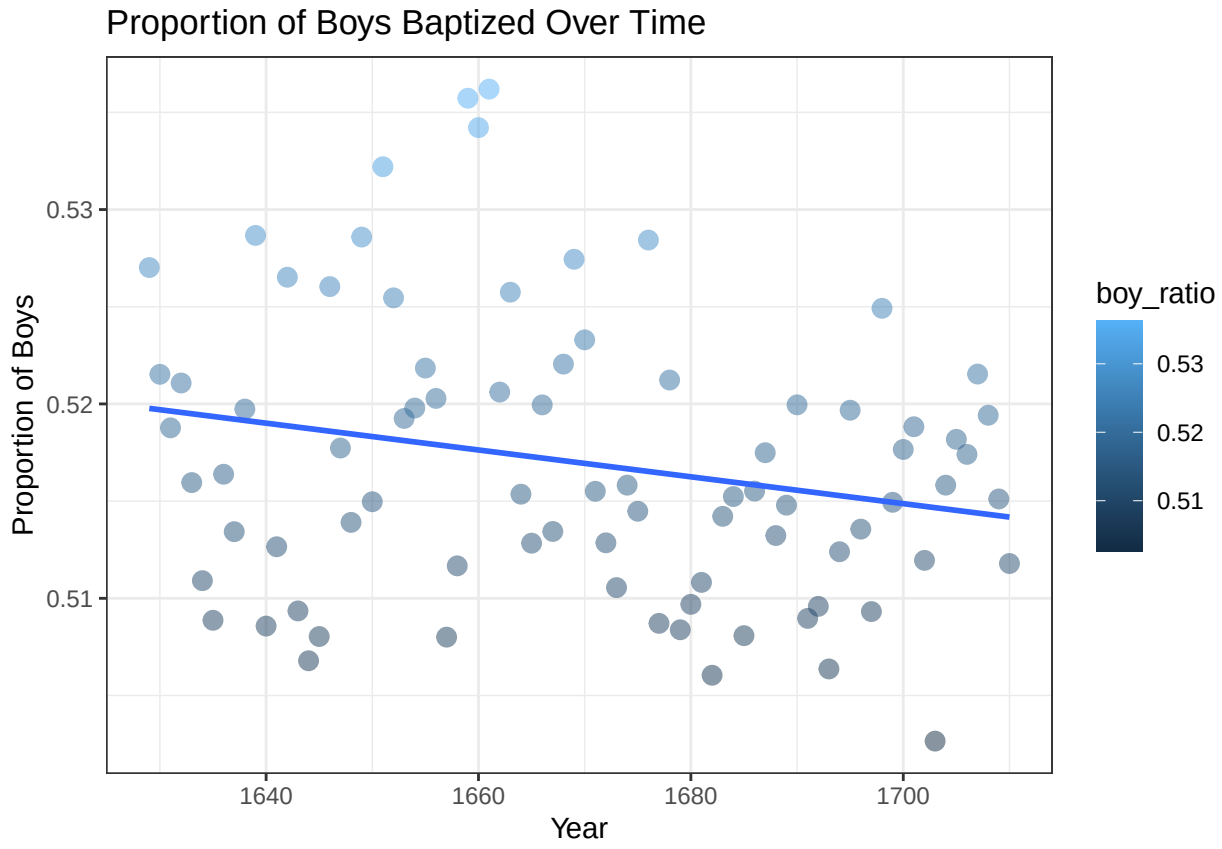
Boys baptized rate decrease.

```
# Insert code for Exercise 3 here
arbuthnot <- arbuthnot %>%
  mutate(boy_to_girl_ratio = boys / girls)
arbuthnot <- arbuthnot %>%
  mutate(total = boys + girls)
arbuthnot <- arbuthnot %>%
  mutate(boy_ratio = boys / total)

arbuthnot <- arbuthnot %>%
  mutate(boy_ratio=boys/total)
  ggplot(arbuthnot, aes(x= year,
                        y= boy_ratio,
                        colour = boy_ratio))+
  geom_point(size= 3, alpha= 0.5)+
  geom_smooth(method= lm,
              se=F)+
  labs(title="Proportion of Boys Baptized Over Time",
       x="Year",
       y="Proportion of Boys")+
  theme_bw()
```

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

## Warning: The following aesthetics were dropped during statistical transformation:
## colour.
## i This can happen when ggplot fails to infer the correct grouping structure in
##   the data.
## i Did you forget to specify a `group` aesthetic or to convert a numerical
##   variable into a factor?
```



Exercise 4

The years included in this data set is year 1940- 2002. The data frame is num. The variable's name are year, boys, and girls.

```
# Insert code for Exercise 4 here
data('present', package='openintro')
str(present)
```

```
## tibble [63 x 3] (S3: tbl_df/tbl/data.frame)
##  $ year : num [1:63] 1940 1941 1942 1943 1944 ...
##  $ boys : num [1:63] 1211684 1289734 1444365 1508959 1435301 ...
##  $ girls: num [1:63] 1148715 1223693 1364631 1427901 1359499 ...
```

```
summary(present)
```

```
##       year           boys             girls
```

```
##  Min.   :1940   Min.   :1211684   Min.   :1148715
##  1st Qu.:1956   1st Qu.:1799857   1st Qu.:1711405
##  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 5

The number of birth records is more then number of baptism. The Present data set is much bigger than Arbuthnot's data

```
# Insert code for Exercise 5 here
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~
```

```
glimpse(arbuthnot)
```

```
## Rows: 82
## Columns: 6
## $ year              <int> 1629, 1630, 1631, 1632, 1633, 1634, 1635, 1636, 1637~
## $ boys              <int> 5218, 4858, 4422, 4994, 5158, 5035, 5106, 4917, 4703~
## $ girls             <int> 4683, 4457, 4102, 4590, 4839, 4820, 4928, 4605, 4457~
## $ boy_to_girl_ratio <dbl> 1.114243, 1.089971, 1.078011, 1.088017, 1.065923, 1.~
## $ total             <int> 9901, 9315, 8524, 9584, 9997, 9855, 10034, 9522, 916~
## $ boy_ratio         <dbl> 0.5270175, 0.5215244, 0.5187705, 0.5210768, 0.515954~
```

Exercise 6

Boy's birth decrease over the time. The Arbuthnot's observation about boys being born in greater proportion than girls does not hold up in the U.S.

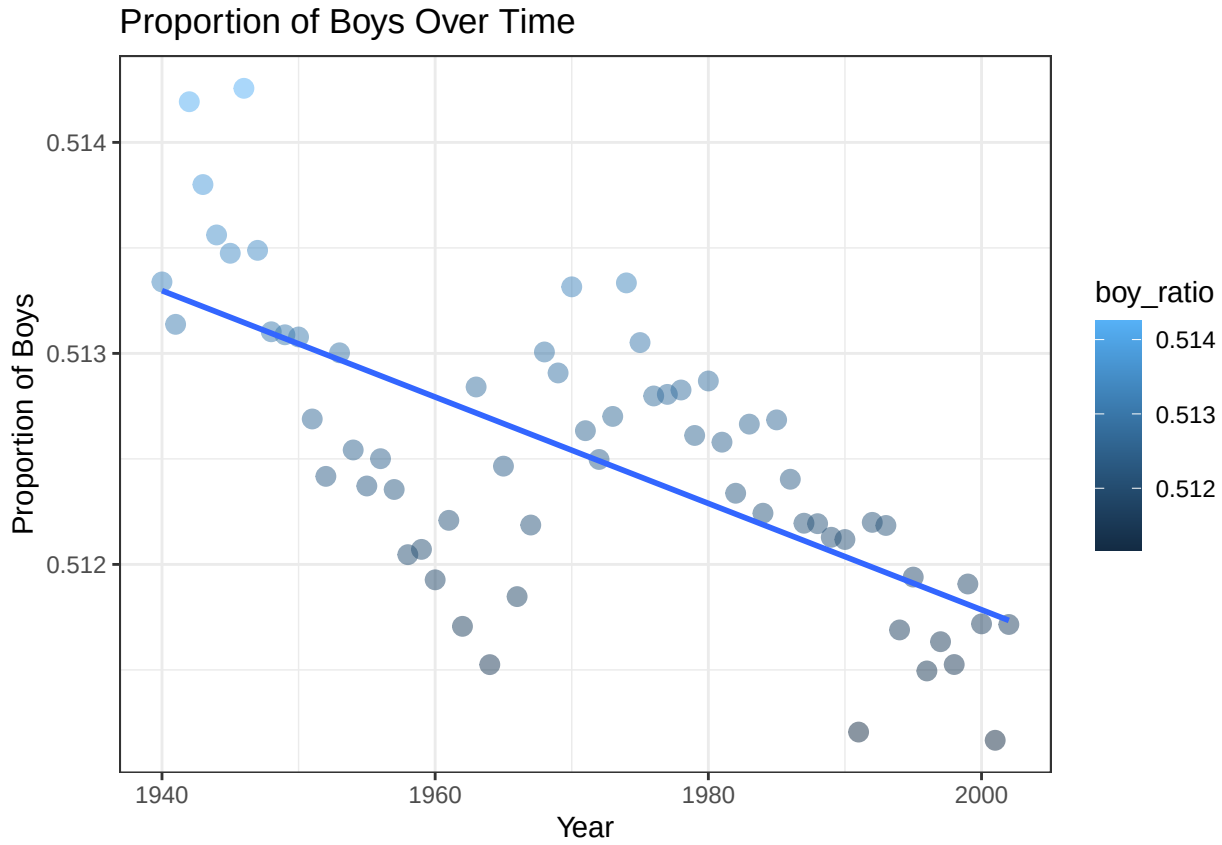
```
# Insert code for Exercise 6 here
present<- present %>%
  mutate(boy_to_girl_ratio = boys / girls)
present <- present %>%
  mutate(total = boys + girls)
present <- present %>%
  mutate(boy_ratio = boys / total)

present <- present %>%
  mutate(boy_ratio=boys/total)
  ggplot(present, aes(x= year,
                      y= boy_ratio,
                      colour = boy_ratio))+
  geom_point(size= 3, alpha= 0.5)+
```

```
  geom_smooth(method= lm,
              se=F)+
  labs(title="Proportion of Boys Over Time",
       x="Year",
       y="Proportion of Boys")+
  theme_bw()
```

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

```
## Warning: The following aesthetics were dropped during statistical transformation:
## colour.
## i This can happen when ggplot fails to infer the correct grouping structure in
##   the data.
## i Did you forget to specify a `group` aesthetic or to convert a numerical
##   variable into a factor?
```



Exercise 7

Year 1961 has most total number of births in the U.S.

```
# Insert code for Exercise 7 here
present %>%
  arrange(desc(total))
```

```
## # A tibble: 63 x 6
```

```
##     year    boys   girls boy_to_girl_ratio   total boy_ratio
##    <dbl>   <dbl>   <dbl>             <dbl>   <dbl>     <dbl>
##  1  1961 2186274 2082052              1.05 4268326     0.512
##  2  1960 2179708 2078142              1.05 4257850     0.512
##  3  1957 2179960 2074824              1.05 4254784     0.512
##  4  1959 2173638 2071158              1.05 4244796     0.512
##  5  1958 2152546 2051266              1.05 4203812     0.512
##  6  1962 2132466 2034896              1.05 4167362     0.512
##  7  1956 2133588 2029502              1.05 4163090     0.513
##  8  1990 2129495 2028717              1.05 4158212     0.512
##  9  1991 2101518 2009389              1.05 4110907     0.511
## 10  1963 2101632 1996388              1.05 4098020     0.513
## # i 53 more rows
```