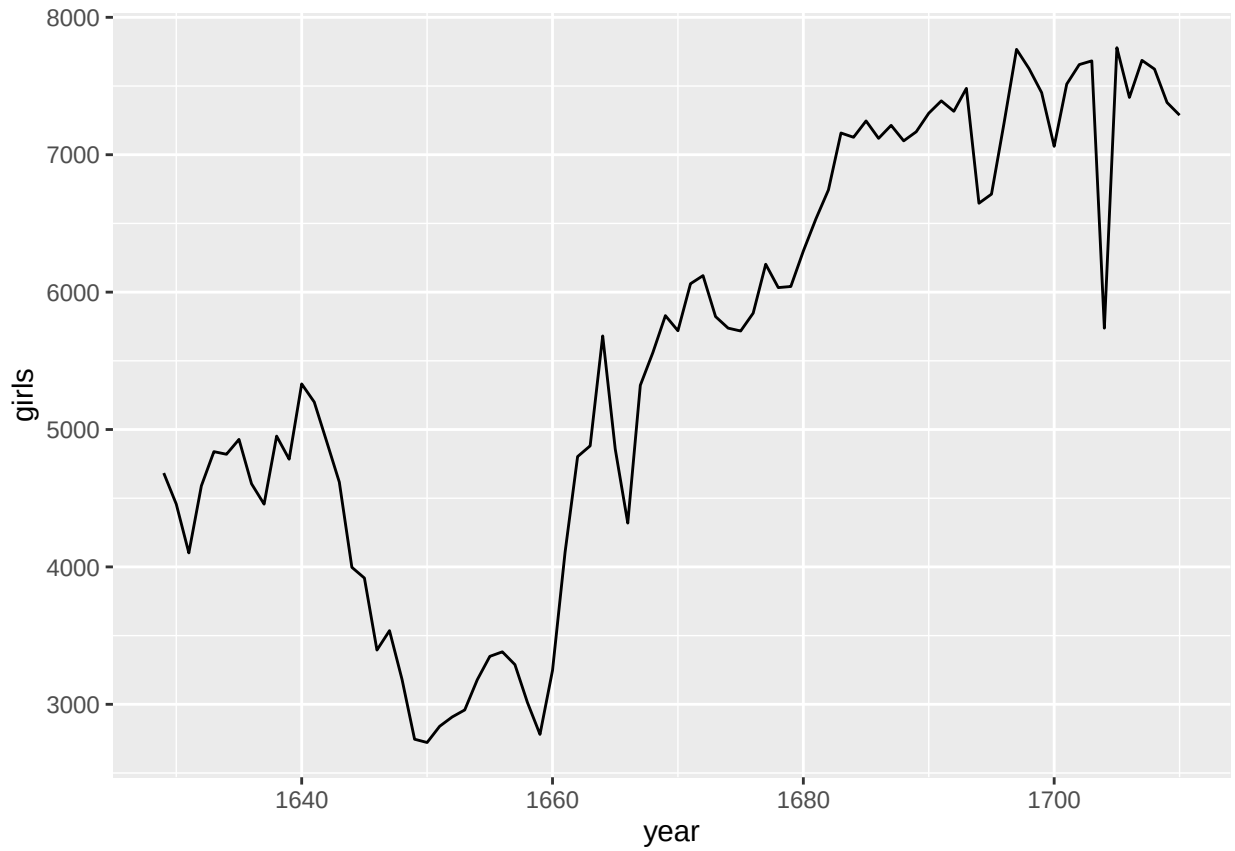
Lab 2 - Exploratory Data Analysis Part I Solution

MATH224 - Intro to Stat

Exercise 1

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



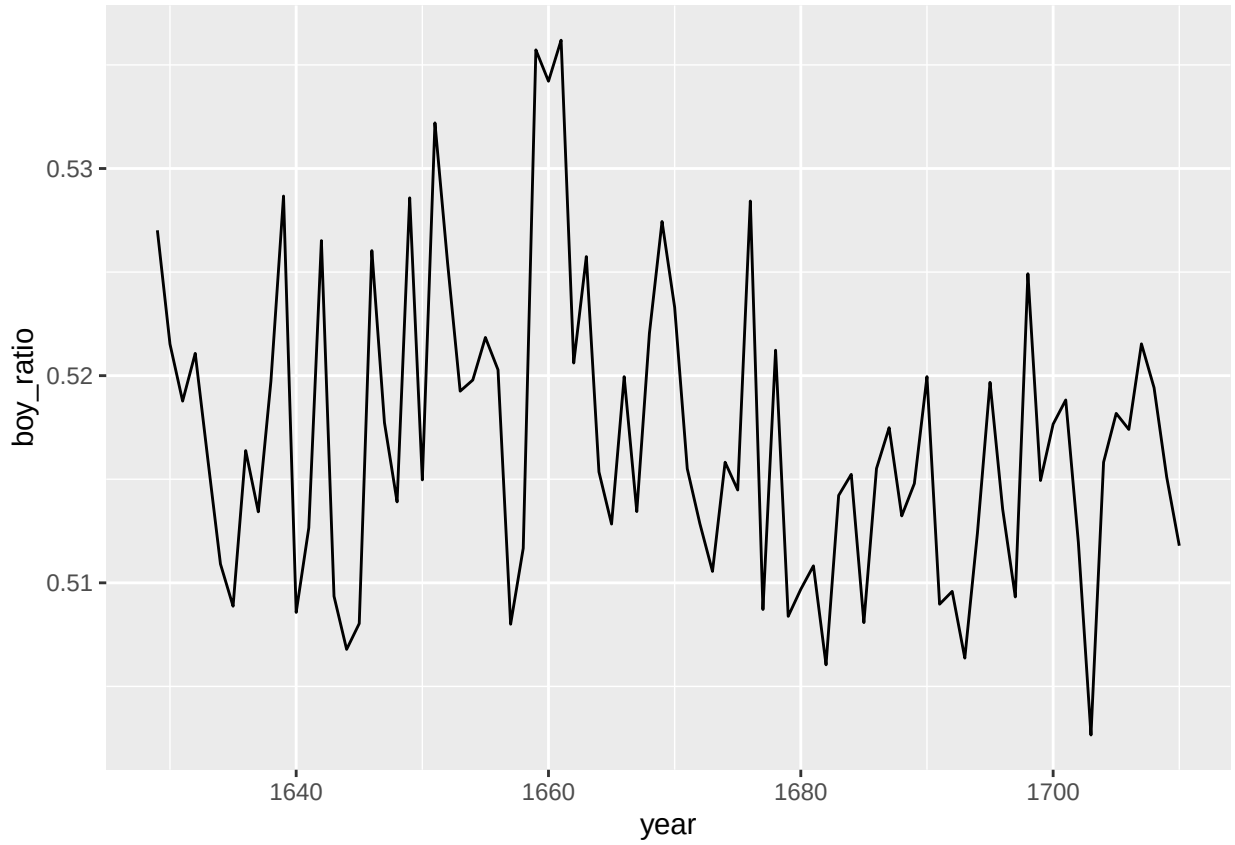
In general, there seems to be an increasing trend throughout the years. We see the least number of girls being born around the year of 1650 followed by 1660. We can also notice that before 1670 the number of girls being born stayed below 6000 and after 1670 it stayed well above 6000.

Exercise 2

```r
# 12 Points
arbuthnot <- arbuthnot %>%
  mutate(total = boys + girls)

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

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



The plot doesn't seems to have apparent trend like the previous exercise. The proportions are all over the place throughout the years but the proportion stays over 0.50 which means that boys were born more than girls in all years.

More Practice

Exercise 3

```r
fivenum(present$year)
```

```
## [1] 1940.0 1955.5 1971.0 1986.5 2002.0
```

The dataset includes years from 1940 to 2002.

```
dim(present)
```

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

The dataset has 63 rows (observations) and 3 columns (variables).

```
names(present)
```

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

Exercise 4

```
glimpse(arbuthnot)
```

```
## Rows: 82
## Columns: 5
## $ year      <int> 1629, 1630, 1631, 1632, 1633, 1634, 1635, 1636, 1637, 1638, ~
## $ boys      <int> 5218, 4858, 4422, 4994, 5158, 5035, 5106, 4917, 4703, 5359, ~
## $ girls     <int> 4683, 4457, 4102, 4590, 4839, 4820, 4928, 4605, 4457, 4952, ~
## $ total     <int> 9901, 9315, 8524, 9584, 9997, 9855, 10034, 9522, 9160, 10311~
## $ boy_ratio <dbl> 0.5270175, 0.5215244, 0.5187705, 0.5210768, 0.5159548, 0.510~
```

```
glimpse(present)
```

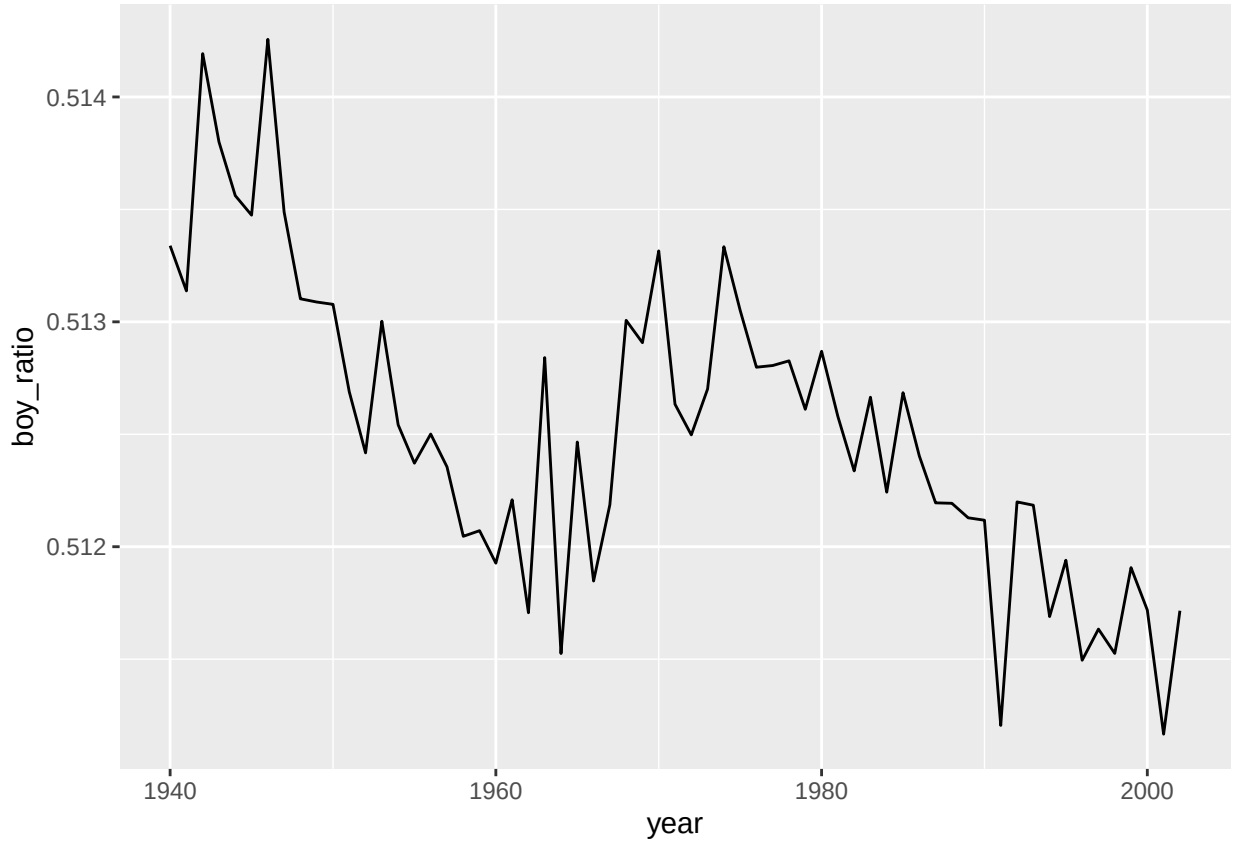
```
## Rows: 63
## Columns: 3
## $ year  <dbl> 1940, 1941, 1942, 1943, 1944, 1945, 1946, 1947, 1948, 1949, 1950~
## $ boys  <dbl> 1211684, 1289734, 1444365, 1508959, 1435301, 1404587, 1691220, 1~
## $ girls <dbl> 1148715, 1223693, 1364631, 1427901, 1359499, 1330869, 1597452, 1~
```

From the output above, we can see that both datasets have differing counts for boys and girls. This could be possibly because of data collection startegies. As we are considering two different countries and two different time periods.

Exercie 5

```
present = present %>%
  mutate(total = boys + girls,
         boy_ratio = boys/total)

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



This plot is completely different from the plot we obtained in exercise 2. Here, we see a clear downward trend in the boys ratio as year increases. But the boy ratio is still above 0.5 meaning that there are more boys than girls being born.

```
present %>%
  arrange(desc(total))%>%
  head(5)
```

```
## # A tibble: 5 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
```

We see that the most total number of births in the U.S. happened in the year if 1961, followed by 1960.