EDA: Present birth records

2022-07-19

loading the R Packages

```
library(dplyr)
library(ggplot2)
library(statsr)
```

Dataset 1: Present birth records

The dataset is regarding the present day birth records in the United States

```
data(present)
```

How many variables are included in this data set?

```
names(present)
```

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

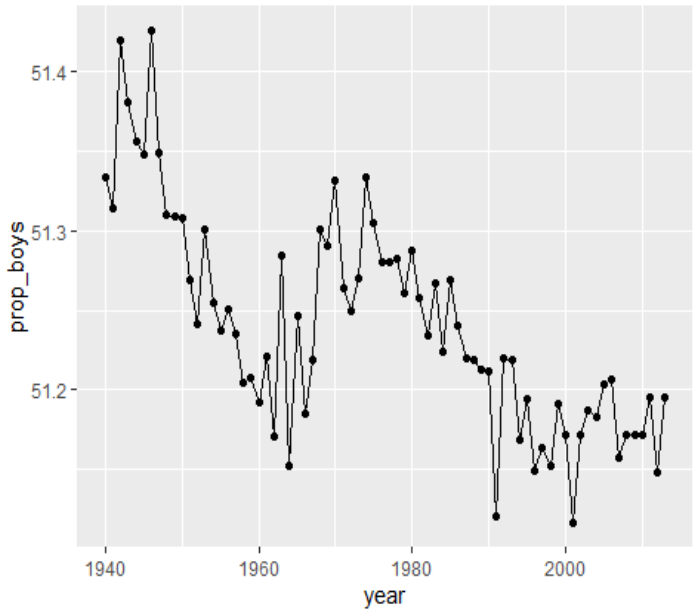
What years are included in this dataset?

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

```
## [1] 1940 2013
```

Calculate the total number of births for each year and store these values in a new variable called total in the present dataset. Then, calculate the proportion of boys born each year and store these values in a new variable called prop_boys in the same dataset. Plot these values over time and based on the plot determine if the following statement is true or false: The proportion of boys born in the US has decreased over time.

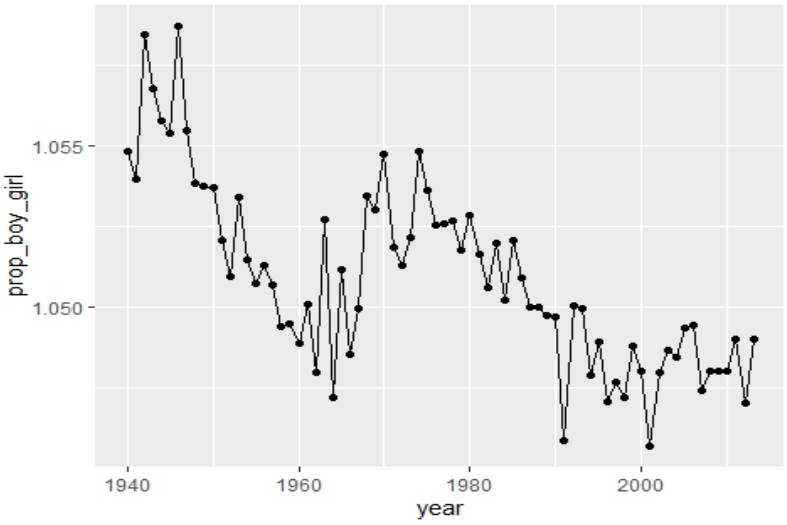
```
present <- present %>%
  mutate(total = boys + girls) %>%
  mutate(prop_boys = (boys/total) * 100 )
  ggplot(data = present, aes(x = year, y = prop_boys)) +
  geom_point()+ geom_line()
```



Calculate the boy-to-girl ratio each year, and store these values in a new variable called prop_boy_girl in the present dataset. Plot these values over time. Which of the following best describes the trend?

1. There appears to be no trend in the boy-to-girl ratio from 1940 to 2013.
2. There is initially an increase in boy-to-girl ratio, which peaks around 1960. After 1960 there is a decrease in the boy-to-girl ratio, but the number begins to increase in the mid 1970s.
3. There is initially a decrease in the boy-to-girl ratio, and then an increase between 1960 and 1970, followed by a decrease.
4. The boy-to-girl ratio has increased over time.
5. There is an initial decrease in the boy-to-girl ratio born but this number appears to level around 1960 and remain constant since then

```
present <- present %>%
  mutate(prop_boy_girl = boys/girls)
  ggplot(data = present, aes(x = year, y = prop_boy_girl)) +
  geom_point()+ geom_line()
```



In what year did we see the most total number of births in the U.S.? Hint: Sort your dataset in descending order based on the total column. You can do this interactively in the data viewer by clicking on the arrows next to the variable names. Or to arrange the data in a descenting order with new function: descr (for descending order).

1. 1940
2. 1957
3. 1961
4. 1991
5. 2007

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