US_Birth_Records_Analysis.R

Bibob

Mon Jul 23 12:23:42 2018

```
# Author:   Bibobra Alabrah

# title:    United States of America Birth Records Analysis

# Dataset:  US Present Birth Records

# Date:     6/7/2018
```

```
# Import dependencies(libraries)
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 3.5.1
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
library(statsr)

# Load the dataset:
data(present)

# View the number of records and features of the dataset:
dim(present)
```

```
## [1] 74  3
```

```r
# Rename the dataset to birth_records:

birth_records <- present
rm(present)

# View the first 10 records of the dataset
head(birth_records, 10)
```

```
## # A tibble: 10 x 3
##     year    boys   girls
##    <dbl>   <dbl>   <dbl>
##  1  1940 1211684 1148715
##  2  1941 1289734 1223693
##  3  1942 1444365 1364631
##  4  1943 1508959 1427901
##  5  1944 1435301 1359499
##  6  1945 1404587 1330869
##  7  1946 1691220 1597452
##  8  1947 1899876 1800064
##  9  1948 1813852 1721216
## 10  1949 1826352 1733177
```

```r
# What years are included in this dataset?
range(birth_records$year)
```

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

```r
# We see that the birth records span from 1940 to 2013.

# What is the total number of births for each year?

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

```
## # A tibble: 6 x 4
##    year    boys   girls   total
##   <dbl>   <dbl>   <dbl>   <dbl>
## 1  1940 1211684 1148715 2360399
## 2  1941 1289734 1223693 2513427
## 3  1942 1444365 1364631 2808996
## 4  1943 1508959 1427901 2936860
## 5  1944 1435301 1359499 2794800
## 6  1945 1404587 1330869 2735456
```

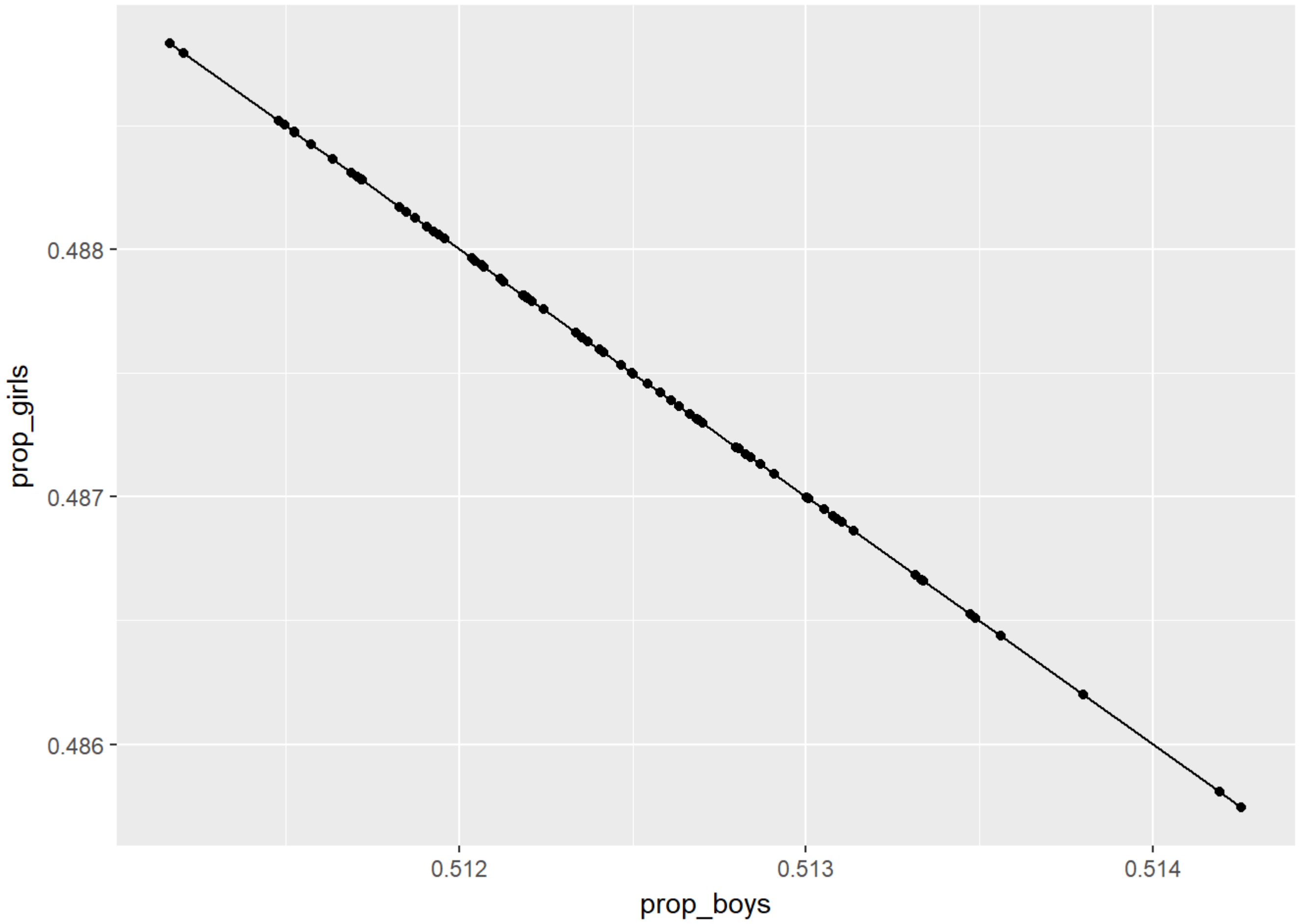
```
# What is the proportion of boys born each year?
birth_records <- birth_records %>%
  mutate(prop_boys = boys/total)
head(birth_records$prop_boys)
```

```
## [1] 0.5133386 0.5131376 0.5141926 0.5138001 0.5135613 0.5134745
```

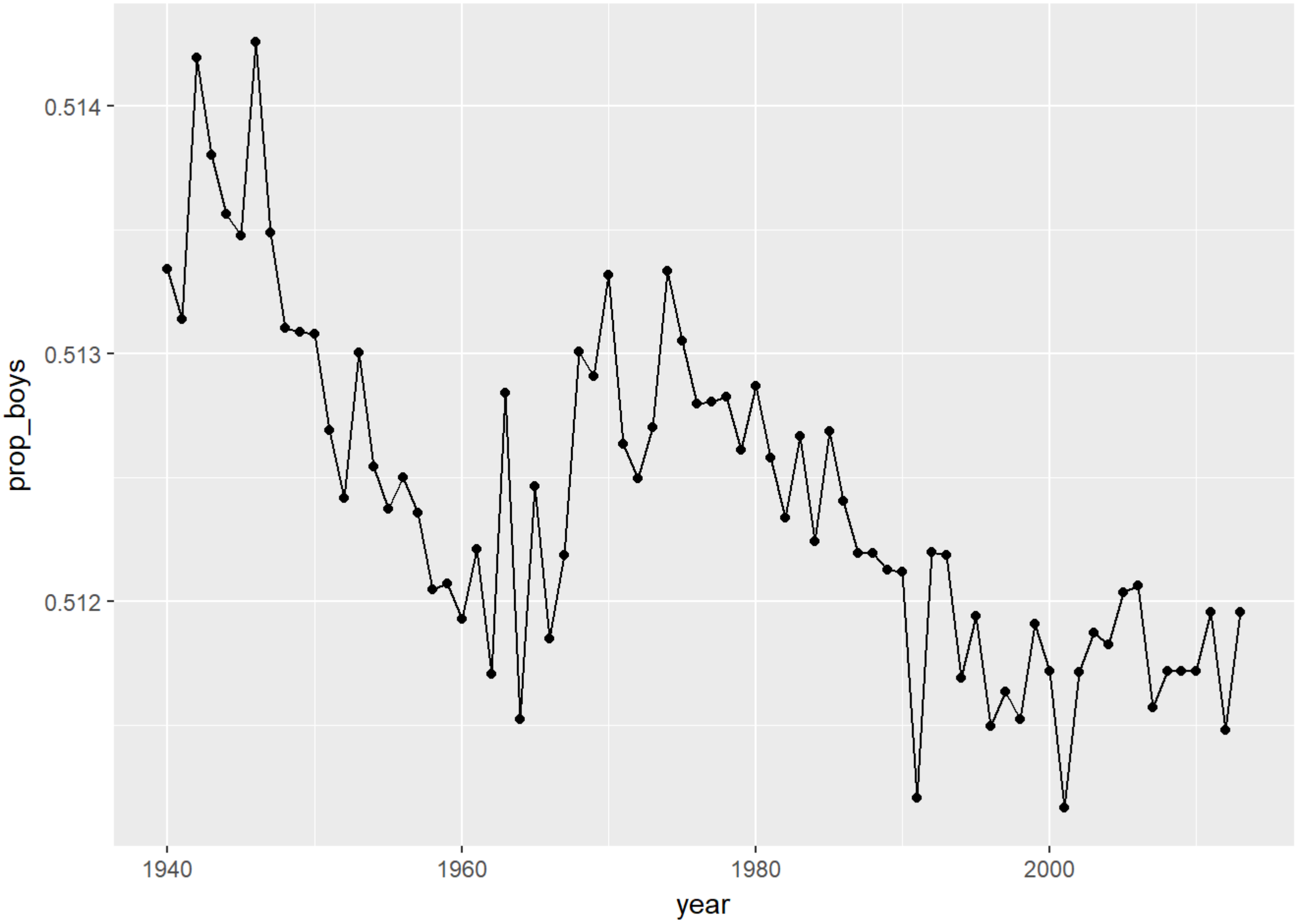
```
# What about the proportion of girls?
birth_records <- birth_records %>%
  mutate(prop_girls = girls/total)
head(birth_records$prop_girls)
```

```
## [1] 0.4866614 0.4868624 0.4858074 0.4861999 0.4864387 0.4865255
```

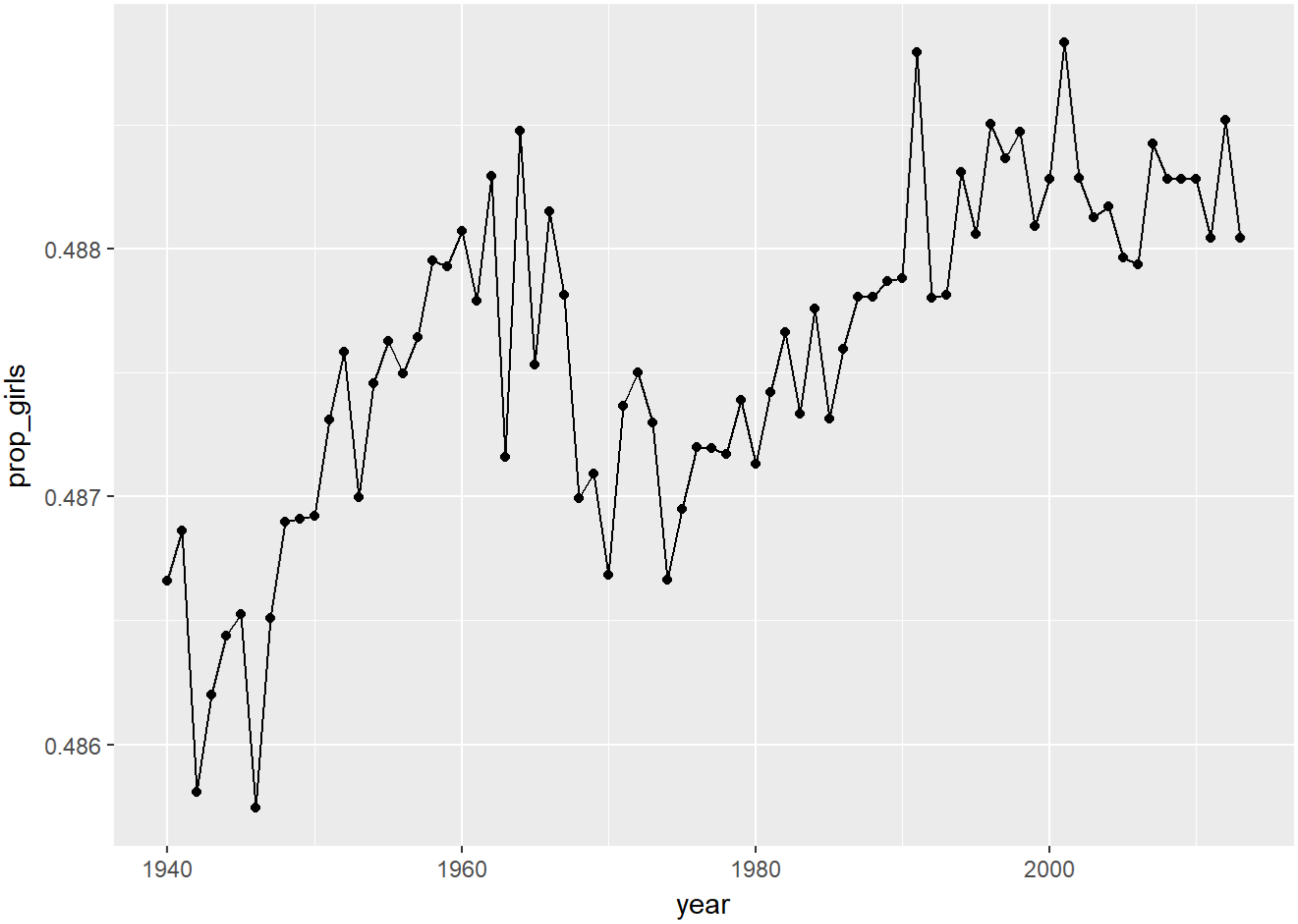
```
# We can see that generally more boys were born during this time.
# Let us visualize this information.

ggplot(birth_records, aes(x = prop_boys, y = prop_girls)) +
  geom_point()+
  geom_line()
```



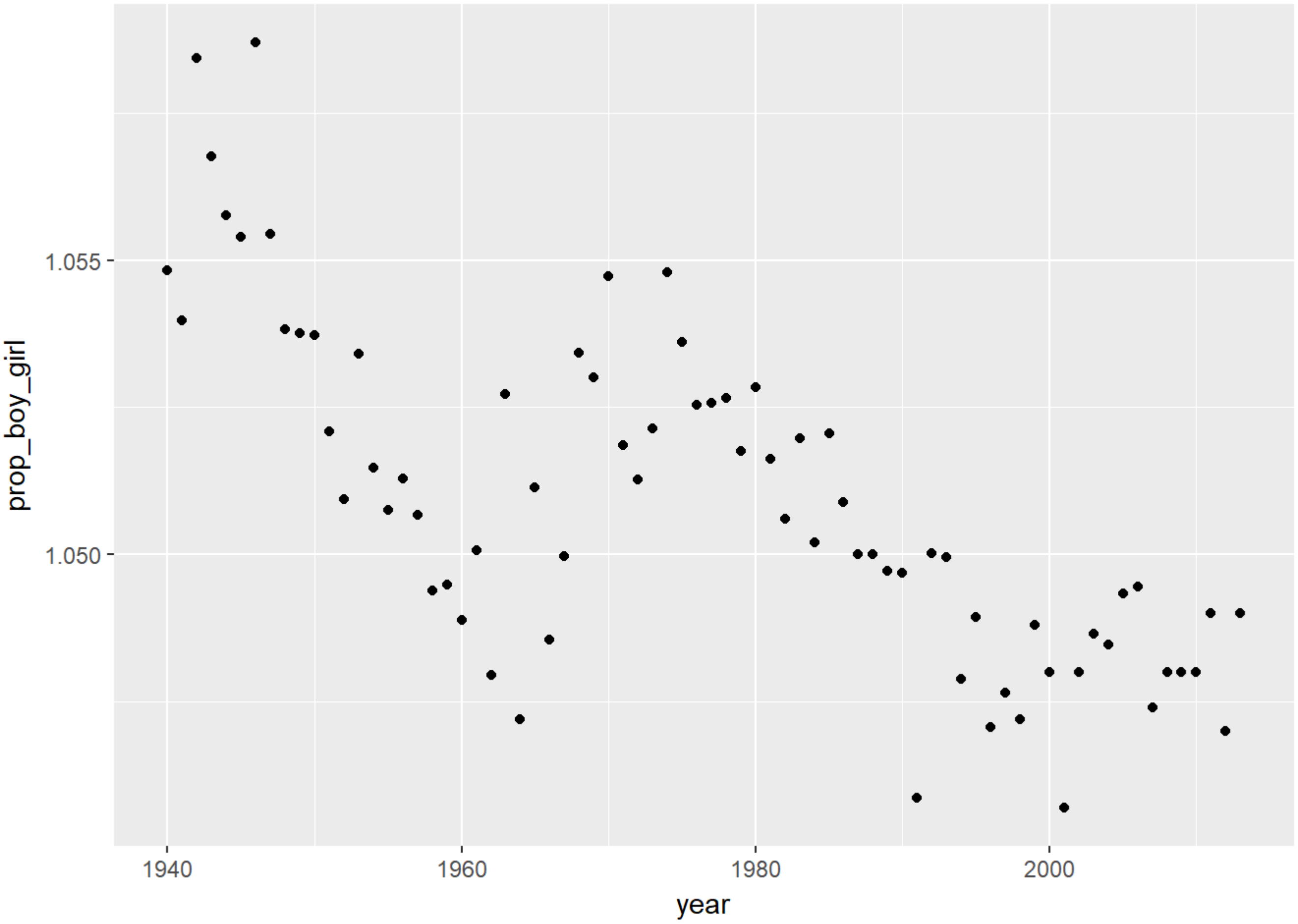
```
# Let us Plot these values over time and based on the plot determine if the
# following statement is true or false: The proportion of boys born in the US has
# decreased over time.

ggplot(birth_records, aes(x = year, y = prop_boys)) +
  geom_point()+
  geom_line()
```



```
# Based on the plot we can see that the proportion of boys born in the US has
# decreased over time.

# Has the number of girls increased over time?
ggplot(birth_records, aes(x = year, y = prop_girls)) +
  geom_point()+
  geom_line()
```



```
# Based on the information, the number of girls has increased over time.

# What is the boy to girl ratio for each?
birth_records <- birth_records %>%
  mutate(prop_boy_girl = boys/girls)

# Plot these values over time. Describe the observed trend?
ggplot(birth_records, aes(x = year, y = prop_boy_girl)) +
  geom_point()
```



```
# There is initially a decrease in the boy-to-girl ratio, and then an increase
# between 1960 and 1970, followed by a decrease.

# In what year did we see the most total number of births in the U.S.?
birth_records %>%
 mutate(total = total) %>%
arrange(desc(total))
```

```
## # A tibble: 74 x 7
##     year    boys   girls   total prop_boys prop_girls prop_boy_girl
##    <dbl>   <dbl>   <dbl>   <dbl>     <dbl>      <dbl>         <dbl>
##  1  2007 2208071 2108162 4316233     0.512      0.488          1.05
##  2  1961 2186274 2082052 4268326     0.512      0.488          1.05
##  3  2006 2184237 2081318 4265555     0.512      0.488          1.05
##  4  1960 2179708 2078142 4257850     0.512      0.488          1.05
##  5  1957 2179960 2074824 4254784     0.512      0.488          1.05
##  6  2008 2173625 2074069 4247694     0.512      0.488          1.05
##  7  1959 2173638 2071158 4244796     0.512      0.488          1.05
##  8  1958 2152546 2051266 4203812     0.512      0.488          1.05
##  9  1962 2132466 2034896 4167362     0.512      0.488          1.05
## 10  1956 2133588 2029502 4163090     0.513      0.487          1.05
## # ... with 64 more rows
```

```
# We see that the US had the most total number of births in 2007.
```